ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre 3224

THESE

En vue de l'obtention du : ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications (LRIT)
Discipline : Sciences de l'ingénieur
Spécialité : Informatique et Télécommunications

Présentée et soutenue le : *10/07/2019* ***par :***

Meryem TALHA

Analyse Syntactico-sémantique automatique de la langue Amazighe

JURY

Salma MOULINE	PES	Faculté des Sciences, Université Mohammed V, Rabat	Présidente
Mounir AIT KERROUM	PH	École Nationale de Commerce et de Gestion, Université Ibn-Tofail, Kénitra	Rapporteur/Examinateur
Najlae IDRISSI	PH	Faculté des Sciences et Techniques, Université Sultan Moulay Slimane, Béni Mellal	Rapporteuse
Abdelali LASFAR	PH	Ecole Supérieure de Technologie, Université Mohammed V, Rabat	Rapporteur/ Examinateur
Ahmed HAMMOUCH	PES	Direction de la Recherche Scientifique et de l'Innovation, Institut des Etudes et Recherches pour l'Arabisation, Rabat	Directeur de thèse
Fadoua ATAA ALLAH	CH	IRCAM : Institut Royal de la Culture Amazighe, Rabat	Examinatrice
Siham BOULAKNADEL	CH	IRCAM : Institut Royal de la Culture Amazighe, Rabat	Encadrante

Année Universitaire: 2019

Dédicace

A

Mes parents

Mon adorable sœur

Mes frères

Toute ma famille, mon beau-frère, mes neveux, mes nièces

Tous mes amis

Tous les Khouribguis

Et à tous ceux et toutes celles qui m'ont accompagné et soutenu durant ces années

Remerciement

Les travaux présentés dans ce mémoire ont été effectués au sein du Laboratoire de Recherche en Informatique et Télécommunications (LRIT-Unité associé au CNRST) à la Faculté des Sciences de Rabat (FSR), Université Mohammed V au Maroc, sous la direction du Feu Monsieur Driss ABOUTAJDINE, ensuite sous la direction de Monsieur Ahmed HAMMOUCH et avec l'encadrement de Madame Siham BOULAKNADEL.

Mes sentiments de reconnaissance vont en premier lieu à mon directeur de thèse Mr. Driss ABOUTAJDINE, ancien Directeur du Centre National pour la Recherche Scientifique et Technique (CNRST) et ancien Responsable du Laboratoire de Recherche LRIT, que Dieu bénisse son âme. D'une part, il a accepté de diriger mes travaux de recherche et m'intégrer au sein du laboratoire LRIT. D'une autre part, Mr. ABOUTAJDINE n'a pas hésité à me soutenir malgré ses nombreuses occupations.

Je tiens à exprimer ma profonde gratitude à Mr. Ahmed HAMMOUCH, Professeur d'enseignement supérieur à l'École Normale Supérieure de l'Enseignement Technique (ENSET) et ancien directeur du Centre National pour la Recherche Scientifique et Technique, pour avoir accepté d'être mon directeur de thèse après le décès de Mr. Driss ABOUTAJDINE. Je le remercie pour sa compréhension, sa gentillesse et son temps précieux qu'il m'a accordé afin de pouvoir soutenir ma mémoire.

Je tiens particulièrement à exprimer mes plus vifs remerciements à mon co-encadrante Madame Siham BOULAKNADEL, Chercheure Habilitée à l'Institut Royal de la Culture Amazighe (IRCAM), pour sa disponibilité, les années de soutien, ses précieux conseils, son attention aiguë aux tous petits détails, sa confiance et surtout la grande rigueur dont vous avez fait preuve. Qu'elle en soit infiniment remerciée !

J'exprime toute ma reconnaissance à Madame Salma MOULINE, Professeur d'enseignement supérieur à la Faculté des Sciences de Rabat, pour avoir bien voulu accepter de présider le jury de ce mémoire.

Monsieur Mounir Ait KERROUM, Professeur Habilité à l'Université Ibn Toufail de Kénitra, pour l'honneur qu'il m'a fait pour sa participation à mon jury de thèse en qualité de

rapporteur de mon travail, pour le temps consacré à la lecture de ce mémoire, et pour les suggestions et les remarques judicieuses qu'il m'a indiquées.

Monsieur Abdelali LASFAR, Professeur Habilité à École Supérieure de Technologie de Salé, pour sa participation à mon jury de thèse en qualité de rapporteur de mon travail et pour toutes remarques intéressantes qu'il m'a faites.

Madame Najlae IDRISSI, Professeur Habilité à la Faculté des Sciences et Techniques de Béni Mellal, qui a bien voulu juger une grande partie de ce travail en tant que rapporteur. Je la remercie pour le temps consacré à la lecture de ce travail ainsi que pour les commentaires m'ayant permis de l'améliorer.

Que Madame Fadoua ATAA ALLAH, Chercheur Habilité à l'Institut Royal de la Culture Amazighe, trouve ici l'expression de mes vifs remerciements pour m'avoir fait l'honneur de participer au Jury de soutenance.

Merci à mes parents, pour les sacrifices consentis depuis toujours, aucun mot ne pourrait exprimer à leur juste valeur la gratitude et l'amour que je vous porte. Ce mémoire vous est dédié à 200%. Merci à mes frères, mes amis et amiespour leur présence dans ma vie, leur soutien spirituel, émotionnel et affectif inconditionnel surtout tout au long de l'épreuve douloureuse que j'ai traversé.

Je souhaite remercier particulièrement mon adorable sœur « Fatiha », merci mille fois pour l'hébergement, les uniques repas quotidiens que tu m'offrais gracieusement pendant toute cette période.

Pensée à tous les professeurs ayant participé à mon apprentissage depuis le primaire. Que tous ceux qui ont contribué à faciliter mes études trouvent ici le témoignage de ma sincère gratitudeet profonde estime qui viennent du fond de mon cœur.

Merci à toutes et à tous.

Résumé

Face à l'afflux de données textuelles en langue amazighe et à leur diversité, il est nécessaire de mettre au point des systèmes adéquats pour y rechercher des informations. La reconnaissance des Entités Nommées (REN) en langue amazighe s'avère un prétraitement éventuellement essentiel pour de nombreuses applications du traitement automatique des langues (TAL).

Dans ce mémoire de thèse, nous présentons une chaîne de reconnaissance des entités nommées (personnes, lieux, organisations, dates, expressions numériques, monnaies et pourcentages) en amazighe fondée sur une étude synthétique des spécificités de la langue et des entités nommées en amazighe. Dans cette perspective, nous avons d'abord construit manuellement un corpus que nous avons nommé « AMCorp ». Ce corpus est utilisé pour évaluer les résultats de nos systèmes. Les évaluations tenues ont été étaient basées sur les métriques classiques, qui sont : la précision, le rappel et la F-mesure. Trois axes de recherche ont par conséquent été investis durant ce travail de thèse : En première partie, nous avons proposé un système basé sur une approche à base de règles, fondée sur un ensemble de règles linguistiques et un ensemble de lexiques créés manuellement. Cette approche a donné des résultats promoteurs, mais en ce qui concerne les textes amazighs elle reste restreinte.

En une seconde étape, nous avons proposé un système de reconnaissance basé sur une approche d'apprentissage automatique supervisé, nous avons utilisé le SVM comme classifieur ainsi qu'un ensemble de caractéristiques de mots amazighes. Les résultats obtenus montrent qu'il est ainsi possible de reconnaitre les entités nommées amazighes avec une très bonne f-mesure.

Les résultats encourageants sur la première et la deuxième approche nous ont poussé à combiner les deux approches précédentes afin d'améliorer le taux de reconnaissance des entités nommées amazighes. La méthode donne des résultats satisfaisants et surpasse la performance des deux autres précédentes.

Mots clés : *Langue amazighe, Traitement automatique des langues naturelles, Reconnaissance des Entités Nommées, Approches à base de règles, Approches à base d'apprentissage, Approches hybrides, Analyse syntaxique, Analyse sémantique.*

Abstract

As more and more Amazighe textual data becomes available and diverse, there is a need to develop adequate systems to process information. Amazighe Named Entity Recognition (NER) system is a potentially vital pretreatment for so many Natural Language preprocessing (NLP) applications.

In this Ph.D. thesis, we present a system for the recognition of named entities (persons, locations, organizations, dates, numbers, currencies and percentages) in Amazigh language based on a synthetic study of the language-specificity and the Amazighe named entities. In this perspective, we have first manually built a corpus that we named "AMCorp". This corpus is used to evaluate the results of our systems. The evaluations held were based on the classical metrics, which are: precision, recall and F-measure. Three main lines were therefore explored during this thesis: in a first part, we propose a system based on a rule-based approach, which uses a set of linguistic rules and a set of lexicons manually created. This approach shows decent performance but is limited to Amazighe texts processing.

In a second step, we propose a recognition system based on a supervised automatic learning approach, we used the SVM as a classifier as well as a set of Amazigh word characteristics. The results obtained show that it is possible to recognize the entities named Amazigh with a very good f-measure.

The encouraging results obtained from the first and the second approaches led us to combine both approaches in order to improve the recognition accuracy of Amazigh named entities. The method gives satisfactory results and surpasses the performance of two previous ones.

Keywords: *Amazighe language, Natural Language Processing, Named Entity Recognition, Rule-based approaches, Machine-Learning based approaches, Hybrid-based approaches, Syntacticanalysis, Semantic analysis.*

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ACE	Automatic Content Extraction
ANNIE	A Nearly New Information Extraction System
CEISIC	Centre des Études Informatiques et des Systèmes d'Information et de Communication
CoNLL	Conference on Computational Natural Language Learning
CRFs	Conditional Random Fields
DARPA	Defense Advanced Research Projects Agency
EN	Entité nommée
ENAMEX	EntityNamed Expression
GATE	General Architecture for Text Engineering
HCP	Haut-Commissariat au Plan
HMM	Hidden Markov Model
IRCAM	Institut Royal de la Culture Amazighe
JAPE	Java Annotation Pattern Engine
LDC	Linguistic Data Consortium
LOC	Location
MEMM	Maximum Entropy Markov Model
MUC	Message Understanding Conferences
NIST	National Institute of Standards and Technology
NUMEX	Number Expression
NLP	Natural Language Processing
NTIC	Nouvelles Technologie d'Information et de Communication
OCR	Optical Character Recognition
POS	Part Of Speech
RAP	Reconnaissance Automatique de la Parole
REN	Reconnaissance des Entités Nommées

RENAM	Reconnaissance des Entités Nommées Amazighes
SER	Slot Error Rate
SVM	Support Vector Machine
TA	Traduction Automatique
TAL	Traitement Automatique des Langues
TALN	Traitement Automatique du Langage Naturel
TIMEX	Time Expression

Introduction générale

1 Contexte de recherche

De nos jours, l'évolution rapide des technologies de l'information et de la communication et la croissance des volumes de données a ouvert d'autres opportunités de manipulation de l'information. Face à ce constat, les recherches en extraction d'information sont devenues de plus en plus nombreuses, dont le but est de permettre à l'utilisateur une exploitation complète des données. L'extraction d'information représente un domaine scientifique pluridisciplinaire, fédérant des thématiques issues des sciences de l'information, de la linguistique, des statistiques et de l'intelligence artificielle. Elle passe par plusieurs traitements allant de la segmentation des données à la compréhension et à la reconnaissance de thèmes. Dans cette chaîne de prétraitements, la tâche d'extraction d'information bénéficie amplement de l'implication de la Reconnaissance des Entités Nommées (REN), dite aussi tâche d'extraction d'indices porteurs de sens.

La tâche de reconnaissance des entités nommées a connu des progrès considérables sur le plan de la qualité des résultats obtenus ainsi que sur la diversité des outils disponibles. Cette dernière décennie, elle a fait l'objet d'une attention plus soutenue et suscite aujourd'hui un intérêt certain. Elle apparaît en effet comme fondamentale pour diverses applications du Traitement Automatique des Langues (TAL) notamment, les systèmes de recherche « Question-Réponse », l'indexation automatique des documents ou encore la traduction automatique. Malheureusement, l'avancement de la recherche dans ce domaine, était souvent orienté vers les langues à plus grande diffusion comme l'anglais ou le français et ce n'est que récemment que la recherche s'est orientée vers le multilinguisme.

Ce travail de thèse s'inscrit dans ce contexte, il s'agit de doter langue amazighe d'un système de reconnaissance des entités nommées dont le but d'apporter des éléments informationnels supplémentaires au traitement sémantique.

2 Motivations et contributions

L'amazighe marocain connait aujourd'hui une dynamique sans précédent, notamment sa constitutionnalisation qui lui offre des perspectives et des opportunités inédites en termes de reconnaissance, de promotion, de revitalisation, et d'appropriation sociale dans le cadre d'un processus d'institutionnalisation cadré par la loi organique prévue à l'article 5 de la Constitution. Face à ce constat, l'exploitation des technologies du langage s'avère une opportunité pour l'amazighe marocain pour qu'il puisse remplir pleinement ses nouvelles fonctions en tant que langue et culture.

Dans ce contexte, nous avons orienté notre démarche vers un double objectif. Il s'agit dans un premier temps d'examiner la notion de la reconnaissance des entités nommées, pour décrire les difficultés rencontrées lorsqu'il est question de reconnaissance des entités nommées dans textes en amazighe. Ainsi, nous avons développé des ressources notamment :

- Un corpus de textes écrits en amazighe. Ce corpus sera destinéà servir d'autres chercheurs dans le domaine afin de standardiser la recherche sur la reconnaissance des entités nommées en amazighe. Il permettra également aux chercheurs intéressés par la langue amazighe un accès à des données actuellement dispersées ou non disponibles. Grâce à la possibilité d'accès rapide aux données, ce dernier peut être réutilisable dans d'autres domaines comme par exemple la traduction automatique.
- Un autre type de ressources linguistiques est élaboré, qui consiste en un lexique des entités nommées ou encore nommé gazetteers. Ce lexique est construit manuellement en exploitant des ressources textuelles (pages Web, corpus, etc.). Il couvre les entités nommées de type personne, lieu, organisation, expressions numériques et expressions temporelles transcrites en langue amazighe, ainsi qu'un lexique de mots déclencheurs permettant l'extraction des entités nommées en question.

Dans un second temps, nous nous sommes intéressés à développer des systèmes de reconnaissance des entités nommées en amazighe qui se basent sur trois approches différentes, à savoir :

- Un système de reconnaissance des entités nommées en amazighe basé sur une approche orientée connaissances, qui repose sur l'implémentation de règles que l'on suppose efficaces pour la tâche considérée. Néanmoins, elle n'est pas applicable à tous les textes en amazighe, le lexique et les règles faites à la main sont souvent optimisés pour un certain type de textes en l'occurrence des textes journalistiques.

Ceci implique donc un enrichissement du lexique et la réécriture de certaines règles à chaque changement du type de textes (Talha, et al., 2014; Boulaknadel, et al., 2014; Talha, et al., 2015; Talha, et al., 2014).

- Un système de reconnaissance des entités nommées en amazighe fondé sur une approche orientée données, qui repose sur l'utilisation des techniques statistiques pour « apprendre » des régularités sur de larges corpus de textes où les entités cibles ont été préalablement annotées (Talha, et al., 2018).
- Un système de reconnaissance des entités nommées en amazighe exploitant une approche hybride. En ceci, nous nous plaçons dans une perspective intermédiaire entre les systèmes orientés connaissances et les systèmes orientés données (Talha, et al., 2015; Talha, et al., 2018).

3 Organisation de la thèse

La suite du mémoire s'organise en cinq parties et se termine par une conclusion.

Le premier chapitre est dédié à la présentation de la tâche de la reconnaissance des entités nommées, son historique et ses différentes applications, tout en décrivant les campagnes d'évaluation menées sur le sujet.

Le deuxième chapitre est une continuité de l'antécédent, où nous dressons un état de l'art des différentes approches de reconnaissance des entités nommées dans le texte ainsi que les travaux menés sur chacune de ces approches.

Le troisième chapitre aborde la langue amazighe marocaine, sujet de notre recherche, en présentant ses caractéristiques, les outils linguistiques utiles conçus pour son traitement automatique. Nous terminons ce chapitre par une présentation des problèmes liés à la reconnaissance des entités nommées en langue amazighes.

Le quatrième chapitre présente l'implémentation et l'architecture des différents systèmes que nous avons conçus pour répondre à notre problématique ainsi que les différentes ressources linguistiques nécessaires que nous avons construites pour représenter les différentes entités nommées.

Le cinquième chapitre est consacré à l'application de nos systèmes proposés. Nous présentons dans un premier temps notre corpus amazighe intitulé « AMCorp » dédié à la tâche de REN en langue amazighe en utilisant la plateforme GATE. Nous donnons ensuite

des informations détaillées sur la manière dont il est implémenté. Enfin, nous expérimentons et nous fournissons des résultats relatifs aux performances obtenues par nos systèmes sur notre corpus.

Enfin, la dernière partie vient conclure ce manuscrit en rappelant nos propositions. Nous y proposons aussi les différents points restant à traiter relativement à la tâche de reconnaissance des entités nommées en langue amazighe.

4 Publications

Revues internationales :

- Meryem Talha, SihamBoulaknadel, Ahmed Hammouch: A brief Survey on Named Entity Recognition in Amazighe language. International Journal of Scientific & Engineering Research vol. 9, Issue 8, pp.121-124. 2018.
- Meryem Talha, SihamBoulaknadel, DrissAboutajdine: Development of Amazighe Named Entity Recognition System Using Hybrid Method. Research in Computing Science vol. 90, pp. 151-161. 2015.
- Meryem Talha, Siham Boulaknadel, Driss Aboutajdine: L'apport d'une approche symbolique pour le repérage des entités nommées en langue Amazighe. Revue des Nouvelles Technologies de l'Information vol. RNTI-E-28, pp.29-34. 2015.
- **Article soumis**: Named Entity Recognizer for Resource-Scarce languages: Amazighe language case. The International Arab Journal of Information Technology (IAJIT).

Chapitres de Livre :

- Meryem Talha, SihamBoulaknadel, DrissAboutajdine: Enhancing performance of Hybrid Named Entity Recognition for Amazighe Language. Machine Learning Paradigms: Theory and Applications. vol. 801, pp 211-232. 2018.
- Meryem Talha, SihamBoulaknadel, DrissAboutajdine: Performance Evaluation of SVM Based Amazighe Named Entity Recognition. Advances in Intelligent Systems and Computing. vol. 723, pp: 232-241. 2018.

Communications dans des congrès internationaux avec actes et comité de lecture :

- SihamBoulaknadel, Meryem Talha, DrissAboutajdine: Amazighe Named Entity Recognition using a rule-based approach. pp: 478-484.AICCSA 2014.

- Meryem Talha, Siham Boulaknadel et Driss Aboutajdine: RENAM: Système de Reconnaissance des Entités Nommées Amazighes.pp : 517-524.TALN 2014.
- Meryem Talha, Siham Boulaknadel et Driss Aboutajdine: Système de reconnaissance des entités nommées Amazighes.pp : 629-638. JADT 2014.

Communications dans des congrès nationaux avec comité de lecture :

- Talha Meryem, BoulaknadelSiham, Ahmed Hammouch. "A brief Survey on Named Entity Recognition in Amazighe language". Faculté des Sciences de Kenitra, CITISI :Technology, Innovation & Information System. 14 & 15 Juillet 2018.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Entités Nommées en Amazighe". Séminaire de formation NooJ. Université Ibn Tofail, Labo MISC et IA4NLP, Kénitra, Maroc, du 07 au 11 Novembre 2016.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Vers un système de détection d'Entités Nommées Amazighes". La 3ème édition des Journées Scientifiques URAC n°29 du LRIT, Institut Scientifique – Rabat, le 28 novembre 2015.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Analyse syntactico-sémantique de la langue Amazighe". Journées « DOCTORIALES 2015 », Rabat les 19, 20 et 21 Février 2015.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "RENAM : Système de Reconnaissance des Entités Nommées Amazighes". La 6ème édition des Journées Doctorales en Technologies de l'information et de la CommunicationJDTIC'14, ENSIAS, Rabat – Maroc, les 19 et 20 Juin 2014.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Analyse syntactico-sémantique de la langue Amazighe". Journées « DOCTORIALES 2014 », Rabat les 6,7 et 8 Février 2014.
- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Reconnaissance des entités nommées en Amazighe". La 5èmeédition des Journées Doctorales en Technologies de l'Information et de la Communication JDTIC'13, Faculté des Sciences de l'Université Ibn Tofail à Kénitra, les 29 et 30 Octobre 2013.

- Talha Meryem, Boulaknadel Siham, Aboutajdine Driss. "Analyse syntactico-sémantique de la langue Amazighe". La 2ème édition des Journées Scientifiques URAC n°29 du LRIT, CNRST – Rabat, le 31 mai 2013.

Chapitre 1:
Reconnaissance des Entités Nommées – Etat de l'art

Après avoir introduit notre problématique de recherche, nous allons dans ce chapitre passer en revue la tâche de reconnaissance des entités nommées, y compris un aperçu général sur les principalescampagnes d'évaluation où cette tâche a été introduite.Enfin, nous clôturerons ce chapitre introductif en présentant quelques domaines d'application exploitant la tâche REN.

1. 1 Introduction

Depuis l'apparition de l'informatique, la masse d'information stockée sur des supports numériques est devenu de plus en plus colossale et continue de s'accroître exponentiellement, nous nous sommes trouvés face à une grande masse d'informations complexe et hétérogènes. Compte tenu de cette augmentation marquante du volume d'informations, nous sommes arrivés à une situation absolument paradoxale : jamais il n'y a eu autant d'informations disponibles, cependant il devient de plus en plus laborieux pour les utilisateurs de retrouver précisément ce qu'ils recherchent dans cette grande masse. La tâche devient de plus en plus pénible et constitue un défi pour les systèmes d'extraction d'informations et les moteurs de recherche.

Quand on parle de l'accès à l'information pertinente, on fait appel au grand domaine de « la recherche ou l'extraction d'informations ». C'est en faveur du développement rapide de la tâche d'extraction d'information que la tâche de reconnaissance des Entités Nommées s'est manifestée.

S'agissant de reconnaissance des entités nommées, on désigne éventuellement une analyse syntactico-sémantique, une phase qui vient juste après l'analyse morpho-lexicale lors du traitement d'un texte ou d'un corpus, consiste à attribuer une valeur sémantique à un mot lexical.

Petit à petit, de manière assez remarquée, cette tâche est devenue d'importante utilité pour diverses applications. Dans certains cas, elle peut être définie comme un prétraitement ou un module de renforcement participant à l'amélioration de diverses applications du TAL, dont la tâche REN constitue une composante interne, ou encore peut être définie comme une application indépendante et directe du TAL.

De nombreux travaux se sont focalisés sur la mise en place des systèmes capables, à l'aide des approches automatique ou classique (en se basant sur l'intervention humaine), d'apprendre automatiquement à reconnaître des entités nommées et à les catégoriser. Ces travaux ont traditionnellement été effectués sur des textes issus de divers domaines (terrorisme, gestion de successions), dans des langues (anglais, espagnol, japonais, français, chinois, arabe, etc.) et type de documents (articles de presse, courriers électroniques) (Nadeau, et al., 2007).

1. 2 Aperçu général de la tâche de la Reconnaissance des Entités Nommées

Depuis les débuts du TALN, la compréhension de texte a fait l'objet d'un suivi bien particulier de plusieurs recherches. Étant une discipline de l'informatique très intimement liée avec le traitement automatique des langues, le domaine de la recherche d'information vise à permettre et faciliter l'accès à un ensemble d'informations pertinentes dans une collection de données. Le champ de l'extraction des informations est souvent décomposé en plusieurs sous disciplines qui sont :

- L'extraction d'entités nommées
- L'extraction de descripteurs thématiques (libres ou normalisés)
- L'extraction de phrases importantes sous un point de vue donné
- L'extraction d'attributs
- L'extraction d'associations entre entités nommées et descripteurs
- L'extraction de correspondances multilingues

Souvent présentée comme une sous-discipline de la recherche d'information, l'extraction des entités nommées devient de plus en plus une étape importante dans tout traitement d'extraction d'information. Elle vise à extraire et catégoriser en fonction des types de classes prédéfinis des éléments informationnels à partir d'un texte donné. Ces éléments représentent des unités lexicales particulières, désignant des noms de personnes, des noms d'organisations et des localisations, ensemble auquel sont souvent ajoutés d'autres syntagmes comme les

expressions temporelles et celles numériques mais peuvent aussi se rapporter à des notions plus techniques comme les maladies.

Le résultat de cette tâche correspond à l'étiquetage des entités en question, lequel se matérialise le plus souvent via des balises d'annotations (marqueurs) de début et de fin encadrant l'entité et mentionnant sa typologie. Prenant l'exemple de la phrase suivante :

« *Aux Etats-Unis, la police a publié les vidéos de la mort de KeithLamont Scott, ce Noir a été abattu par la police à Charlotte, en Caroline du Nord, provoquant la colère des habitants et plongeant la ville dans une vive agitation. Les images dévoilées ne permettent pas de voir s'il était armé.* »Extrait du journal le Monde.

Le but est de détecter les entités **Etats-Unis, Keith Lamont Scott**, **Charlotte**, etc., ensuite de leur accorder la classe sémantique convenable. Les systèmes d'annotation en entités nommées existants utilisent la représentation par balise (SGML, XML, etc.) pour étiqueter un texte. L'annotation peut alors se faire de la manière suivante :

« Aux**<loc>Etats-Unis</loc>,** la police a publié les vidéos de la mort de**<pers>Keith Lamont Scott</pers>,** ce Noir a été abattu par la police à**<loc>Charlotte</loc>,** en**<loc>Caroline du Nord</loc>,** provoquant la colère des habitants et plongeant la ville dans une vive agitation. Les images dévoilées ne permettent pas de voir s'il était armé. »

Parfois, on se trouve face à des ambiguïtés sémantiques comme le cas de **Noir** qui, dans cet exemple, réfère à une personne, alors qu'il peut aussi servir à désigner une couleur. Partant de ce constat, on peut déduire que la reconnaissance d'entités nommées ne peut pas se limiter à un traitement de correspondance syntaxique de syntagmes, mais elle demande un traitement sémantique supplémentaire assez complexe.

Dans la grande variété des domaines de recherche en informatique, la Reconnaissance des Entités Nommées (REN) a été initialement et exclusivement déclenchée pour le traitement de textes journalistiques et économiques. Elle a pu, par la suite, recouvrir d'autres champs d'activité comme la biologie ou la microbiologie(Ananiadou, et al., 2007), la médecine (Bodenreider, et al., 2000)ou la chimie (Rocktäschel, et al., 2012). En outre, elle ouvre des voies d'exploration pour les recherches dans différentes disciplines comme la traduction automatique(Babych, et al., 2003), Systèmes Questions-Réponses(Pizzato, et al., 2006), Résumé automatique (Nobata, et al., 2002)et autres.

1. 3 Historique des campagnes d'évaluation

La tâche de reconnaissance d'entités nommées a gagné en maturité et s'est fabuleusement imposée dans le domaine du TAL, grâce à la participation d'un ensemble de campagnes d'évaluation (en anglais evaluationconferences). Ces dernières sont les pionnières dans ce domaine, à travers lesquelles plusieurs évaluations ont été établies. Ce qui a donné l'occasion aux chercheurs de contribuer massivement au progrès de cette tâche. Parmi les objectifs de ces campagnes, était de fournir des données, des outils d'évaluation et mesurer l'efficacité des systèmes pour l'extraction des informations spécifiques et bien définies en général dans des corpus de documents associés.

1. 3. 1 MUC[1] (Message UnderstandingConferences)

Cette série de conférences a commencé aux États-Unis entre 1987 et 1998, et a été organisé par diverses institutions américaines et financé par la DARPA (Defense Advanced ResearchProjects Agency), par la suite des campagnes d'évaluation similaires ont été organisées un peu partout dans le monde. En effet, le but de ces conférences était à l'origine de stimuler et promouvoir la recherche autour de l'analyse automatique des messages militaires contenant des informations textuelles. Cette série, organiséeen 7 éditions qui se sont succédées, a réussi à développer la tâche d'extraction d'information en général, et l'extraction des entités nommées en particulier. Elle a considérablement réactivé ces courants de recherche.

La première édition MUC-1, organisée en 1987, a été une phase préparatoire où aucune nature des informations à traiter n'a été définie au préalable. L'objectif était de faire un état des lieux des systèmes de compréhension de textes et de générer des bases de connaissances. Au cours de cette première édition, les participants se contentaient d'établir le format des sorties de leurs systèmes, aucune implémentation n'a été mise en œuvre. En 1989, vint l'organisation de la deuxième édition MUC-2. Lors de cette dernière, un premier formulaire simple de structuration de données (en anglais Template) a fait son apparition. Le but était de spécifier les informations à extraire, ça représentait un ensemble de champs à remplir contenant des classes sémantiques. En outre, les premières mesures d'évaluation ont été mises en place inspirées de la recherche d'information, à savoir la précision (P) et le rappel (R). La première est le pourcentage des résultats corrects parmi les résultats obtenus, tandis

[1]http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

que la deuxième représente le pourcentage des résultats corrects parmi les résultats qu'on doit trouver.

Les conférences MUC-3 (1991), MUC-4 (1992) et MUC-5 (1993) se différent des deux sessions précédentes principalement par la très grande variété des données à exploiter. Au cours de ces sessions la tâche d'extraction d'information s'est progressivement bien définie, cependant elle est également devenue plus complexe.

Les participants de MUC-3 et MUC-4 ont travaillé sur des corpus de nature journalistique, traitant des évènements et des actes terroristes en Amérique Centrale et du Sud afin d'extraire le maximum d'informations sur des actes terroristes comme le nom de groupes terroristes impliqués, le nom des victimes, les types d'armes utilisées, les dates et les lieux. Les corpus contenaient des textes bien écrits et homogènes ce qui a rendu leur analyse plus difficile. En outre, les formulaires proposés ont subi des modifications, comportent alors plus de champs à remplir allant jusqu'à 24 champs. A partir de là, les systèmes à base d'automates et ceux basés sur des méthodes statistiques ont apparu. En revanche, il s'est avéré, durant MUC-3, que l'utilisation de deux métriques d'évaluation (Précision et Rappel) produit une certaine confusion quant à la comparaison de performances des systèmes, ce qui a permis l'adoption de la F-mesure comme mesure d'évaluation. C'est un compromis entre le rappel et la précision qui rend plus facile les comparaisons entre systèmes. Afin de favoriser la probabilité des systèmes proposés, vint l'édition MUC-5 qui a vu une augmentation de degré de complexité en introduisant deux domaines pour l'extraction d'information pour l'anglais et le japonais. Vu la grande diversité des informations, la réalisation des systèmes adéquats a demandé beaucoup de temps et les niveaux de performance ne dépassaient pas les résultats précédents. Ce qui a poussé les participants à réaliser des architectures génériques avec des modules d'analyse absolument indépendants.

Etant donné la complexité de la tâche d'extraction d'information, et dans l'espoir de répondre aux besoins spécifiques exigés par le comité scientifique et camoufler l'échec des deux dernières éditions, la tâche d'EI a été répartie en 4 tâches indépendantes et complémentaires dans l'édition MUC-6 (1995). Il s'agit de la reconnaissance d'entités nommées, résolution de conférence, désambiguïsation lexicale et détection de structure prédicat-argument. La tâche de reconnaissance d'entités nommées dans MUC-6 consiste à utiliser des marqueurs pour identifier, dans les textes, les noms propres, des expressions temporelles et des expressions numériques, qui sont structurés en trois classes, à savoir :

- **ENAMEX** : Tous les noms de personnes, d'organisation et de lieux.
- **TIMEX** : Les expressions temporelles comme les dates et les temps.
- **NUMEX** : Les expressions numériques comme les monnaies et les pourcentages.

Cette classification exclut l'annotation de certains noms comme les noms de lois, de prix, de maladies, de saints, etc. le bilan de MUC-6 voit se multiplier les méthodes probabilistes et les systèmes à base d'apprentissage et répond aux objectifs fixés préalablement par le comité. Après ce grand succès, la tâche de REN a été également adoptée dans la septième et dernière édition de MUC-7 (1998) (Chinchor, et al., 1997), qui marque en réalité la fin des conférences MUC avec des nouveautés légères par rapport à l'édition précédente. En effet, cette édition s'est distinguée par le fait que les données d'apprentissage et de test ne portaient pas sur le même domaine, (à titre d'exemple : les textes journalistiques rapportant des crashs d'avion et de tirs de missiles).

La série de conférences MUC a fait le succès de la tâche de reconnaissance des entités nommées et a incité les chercheurs et développeurs à concevoir des outils permettant l'extraction de plus d'informations en élargissant le concept d'entité nommée. Pour étendre ce champ de recherche à d'autres domaines et à d'autres langues, d'autres campagnes ont suivi la série MUC.

1. 3. 2 ACE (Automatic Content Extraction)[2]

Le programme ACE (Automatic Content Extraction) a été organisé par le NIST (National Institute of Standards and Technology) entre 2000 et 2004(Doddington, et al., 2004), et a repris le flambeau des campagnes MUC avec la détermination de nouvelles optiques. Il avait pour vision de continuer des travaux sur l'analyse sémantique que les conférences MUC ont introduits. Les participants se concentraient plus sur les traitements sémantiques que vers les traitements linguistiques, en développant les champs de détection des entités nommées, des événements et des relations exprimées entre ces éléments. La compréhension des éléments au sein des entités nommées est devenue indispensable. La vision est devenue différente, une entité fait référence à un concept et non à une chaîne de caractères. Donc, il ne s'agissait plus d'extraire une chaîne de caractères mais bien d'essayer d'extraire le concept d'entité nommée en elle-même. L'entité est donc explorée et annotée dans son intégralité. Elle peut être détectéeà travers ses diverses mentions, à titre d'exemple les noms propres, les expressions nominales ou encore les pronoms.

[2]http://www.itl.nist.gov/iad/mig//tests/ace/

En outre, l'accent est plus centré sur la mise en place des systèmes plus indépendants permettant de couvrir des domaines plus larges et plus robustes comme le traitement des données bruitées provenant de systèmes de Reconnaissance Automatique de la Parole (RAP) ou ayant été numérisées par l'OCR (Optical Character Recognition). Partant de ce constat, l'objectif dans ACE est plus complexe et plus précis.

Les participants ont travaillé sur des corpus d'entraînement et de test contenant des données extrêmement diversifiées (écrites, orales et données bruitées) collectés à partir de sources différentes : blog de discussion sur le web, conversations téléphoniques transcrites, journaux radiophoniques transcrits, etc. avec une distribution d'un tiers de chaque type. Le Linguisitic Data Consortium (LDC)[3] s'occupait de l'évaluation. Le LDC est un consortium de recherche et de développement qui s'intéresse aux technologies liées à la langue et vise à partager des ressources et définir des standards.

Dans la même lancée de MUC, ACE s'est déroulé en 7 éditions, les chercheurs dans la première édition se sont concentrés sur la détection et la classification des entités « EntityDetection and Tracking »(Florian, et al., 2004), comme son nom l'indique. La détection et la classification des entités étaient deux tâches distinctes et ont été évaluées séparément. Dans cette première édition, le comité scientifique a défini quatre secteurs d'activité, à savoir : la détection des entités nommées ; la classification des entités ; la détection des mentions qui permettent la détermination des expressions nominales ou pronominales qui désignent une même entité ; la reconnaissance des extensions des mentions qui repose sur l'extraction des syntagmes décrivant des mentions d'entités.

La typologie définie a subi des modifications importantes et différentes par rapport à la classification introduite aux conférences MUC, en utilisant une catégorisation beaucoup plus pointue et fine des entités nommées. Ainsi 5 classes d'entités ont été définies comme suit :

- **Personnes :** désigne toute personne ou tout groupe de personnes signalé par un nom, un groupe nominal, une fonction ou même un pronom.
- **Organisations :** indique les noms d'organisations ou un ensemble d'organisations ayant une structuration bien définie. Ceci regroupe les entreprises et sociétés, les unités gouvernementales, les équipes sportives, les groupes musicaux... ;

[3]https://www.ldc.upenn.edu/collaborations/past-projects/ace

- **Lieux :** cette classe désigne un lieu défini sur des bases géographiques ou astronomiques et ne constituant pas une entité politique.
- **GSP (Geographical- Social-Politicalentities) :** vise à annoter les entités géographiques définies sur des bases politiques ou sociales : par exemple, dans la phrase « le Maroc s'est imposé, devant le Burkina Faso (2-0) », l'entité nommée « Maroc » réfère en même temps à un lieu et à une organisation.
- **Bâtiments (Facility) :** elle inclut les infrastructures et les bâtiments construits par l'homme comme les habitations, les usines, les stades, les prisons, les musées, les parkings, les routes, les ponts...

La qualité de la REN est évaluée en termes du coût d'appariement Slot Error Rate (SER) (Makhoul, et al., 1999)qui est une adaptation d'ERR (ErrorPer Response).

La deuxième édition ACE (2001-2002) s'est caractérisée par l'introduction d'un nouveau concept qui consiste à détecter les relations existantes entre les entités. Ce concept peut être utile pourdiverses applications en TAL à titre d'exemple la compréhension de textes et l'indexation de documents.

Dans l'édition suivante, ACE (2003) représente une continuité des travaux entamés lors de sa précédente.Les mêmes standards d'annotations et les mêmes métriques d'évaluation ont été conservés, la seule différence, réside dans le fait que la tâche de REN s'est étalée sur deux langues supplémentaires le chinois et l'arabe. Ceci a donné l'occasion aux participants de cette édition de collecter des données pour ces deux langues dans le but de promouvoir le développement d'outils indépendants de la langue permettant de les traiter.

ACE (2004) se distingue par l'insertion de deux nouveaux axes de recherche permettant, d'une part la détection des évènements et des expressions temporelles, et d'autre part, la définition de nouvelles consignes d'annotations, dans le but d'affiner le tri des entités. Deux nouvelles classes ont été insérées à savoir Vehicle (Air, Land, Subarea-Vehicle, Underspecified, Water) et Weapon (Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified).

La campagne d'évaluation ACE (2005) s'est caractérisée par son aspect multilingue. Les chercheurs se sont concentrés sur des données en langue anglaise, chinoise et arabe. En outre, une révision affinée de la typologie des catégories d'entités a été également établie. Ceci implique, que chaque entité possède une catégorie, une sous-catégorie et une classe. De plus, les corpus constitués contenaient des journaux papier, journaux du Web et émissions radio et

télédiffusées, des données de type paroles conversationnelles dans des émissions télévisées, des conversations téléphoniques et des données collectées sur des blogs et des forums de discussion. Tandis qu'ACE (2007) a été marquée par l'implication d'une nouvelle langue, l'espagnol, dans la tâche de REN en plus de l'arabe et du chinois, et par l'introduction d'un nouveau type de recherche qui est celui de la traduction des entités nommées. Partant du fait que les entités sont des mots porteurs de sens, pouvoir traduire ces expressions serait un atout pour améliorer les performances des systèmes multilingues.

ACE (2008) marquait la fin des conférences ACE, n'apportait pas beaucoup de nouveautés, toutefois les métriques d'évaluation pour les tâches de détection d'entités nommées et de détection des relations ont subi de légères modifications. La métrique d'évaluation qui a été utilisée durant cette édition est LEDR-value (Local EntityDetection and Recognition value).

La série ACE a contribué énormément au développement de la tâche de reconnaissance des entités nommées, et bien évidement le traitement sémantique de la langue. Elle a permis ainsi de tracer un contour plus net de la tâche d'extraction des entités nommées, de relations entre elles et des évènements à exploiter dans nombreuses autres campagnes d'évaluation à travers le monde.

1. 3. 3 CoNLL (Computational on Natural Language Learning)[4]

CONLL (Computational on Natural Language Learning) est la conférenceannuelleorganisée par SIGNLL[5] (ACL Special Interest Group on Natural Language Learning). C'est une conférence internationale qui s'intéresse au traitement du langage naturel et plus particulièrement à la tâche de REN, en développant des méthodes fondées sur des approches statistiques. Elle se positionne approximativement de la même manière que MUC. Dans le cadre de cette campagne, il y a une « SharedTask » qui est effectivement une compétition dont l'intention est de permettre aux participants d'évaluer leurs systèmes proposés sur un même jeu de données.

Deux éditions ont été organisées en 2002 et 2003, réunissant plus d'une dizaine de participants et ont travaillé sur des corpus issus de presse. La première a eu lieu en 2002 (Sang, et al., 2002), le sujet principal de cette édition était la reconnaissance des ENs, et

[4]http://www.cnts.ua.ac.be/conll/

http://www.cnts.ua.ac.be/conll2003/ner/

[5]http://www.signll.org/

portait sur le traitement des deux langues l'espagnol et le hollandais. Chacune des langues mentionnées, possède une base d'apprentissage, une base de développement et une de test. Pour la langue espagnole, il s'agit d'une collection d'articles de presse issue de «the Spanish EFE News Agency », datée de Mai 2000. En revanche, le corpus hollandais se compose des articles issus des quatre éditions du journal de la Belgique "De Morgen" de la période 2000 (Juin, Juillet, Aout and Septembre).

Tandis que la deuxième édition (Tjong Kim Sang, et al., 2003)traitait l'anglais et l'allemand. Pour la langue anglaise, Ils utilisaient un corpus journalistique issu de l'agence de presse anglaise REUTERS[6], daté entre août 1996 et août 1997. Quant aux données allemandes, elles sont extraites à partir du « ECI MultilingualText Corpus[7] ». Ce dernier contient des textes de différentes langues, les données extraites pour la langue allemande étaient à partir du journal allemand « FrankfurterRundshau », datées d'août 1992, septembre et décembre 1992.

La difficulté de créer une typologie à large couverture a conduit lors des conférences CoNLL 2002 et 2003 à la mise en place d'une structure hiérarchique assez restreinte des entités nommées hiérarchisée en quatre classes : Personnes, Organisations, Lieux et Divers. Notons que ces classes sont les trois sous-classes de la catégorie ENAMEX de MUC plus une quatrième classe qui regroupe toutes les entités détectées qui ne font pas partie des trois classes précédentes, par exemple, les nationalités (American) et les événements (Jeux Olympiques de 2000). Ce type de classe a été ensuite adopté dans plusieurs projets qui ont suivi.

1. 3. 4 ESTER[8]

La campagne d'évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques (ESTER) a pour objectif de mesurer objectivement les performances des systèmes de transcription d'émissions radiophoniques pour le français et les faire progresser. Cette campagne s'intéresse à la tâche de reconnaissance d'entités nommées pour les journaux et émissions radiophoniques.

Deux éditions ont été organisées dans ce sens, dans le cadre du projet EVALDA (Évaluation des technologies de la langue en français): sous les noms de ESTER 1 (2003-2005) (Meur, et

[6]http://www.reuters.com/

[7]https://catalog.ldc.upenn.edu/LDC94T5

[8]http://www.afcp-parole.org/camp_eval_systemes_transcription/

al., 2004)et ESTER 2 (2006-2008) (Galliano, et al., 2009)**.** Il faut mentionner que dans les deux éditions, les transcriptions sont enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des entités nommées, etc. Et ceci dans l'optique de faciliter la tâche d'extraction d'informations en ayant une transcription lisible et une représentation structurée du document.

La première campagne est financée par le Ministère de la Recherche dans le cadre de l'appel à projet Technolangue[9], sous l'impulsion de l'Association Francophone de la Communication Parlée, du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et d'ELDA[10] (Evaluations and Languageresources Distribution Agency). Cette édition contient un jeu de 30 types d'entités nommées réparties en 9 catégories principales à savoir :

- **Personne :** comprend tous les mots ou expressions désignant des personnes ou animaux, qu'ils soient réels ou fictifs (roman, film, bande dessinée, etc.). Ce groupe de mots répartis en 3 sous-classes à savoir : pers.hum: contenant toutes les personnes réelles; pers.anim: tous les animaux réels; pers.imag: tous les personnages et animaux fictifs.
- **Organisation :** désigne toutes les expressions, noms, sigles, etc. faisant référence à une organisation de nature politique, religieuse, culturelle, etc. Ce type d'entités peut être classé sous 4 sous-classes à savoir : org.pol : les organisations à caractère politique ; org.com : celles à caractère commercial ; org.edu : les organisations à visée éducative etorg.non-profit : les organisations qui sont à but non lucratif.
- **Lieu :**cette catégorie est répartie en 5 sous-catégories :loc.geo: représente les lieux géographique naturels (y compris les fleuves, montagnes, rivières, continents, systèmes solaires, etc.); loc.geo.line : désigne tous les axes de circulation (chemins ferroviaires, tunnels, ponts, rues et boulevards, etc.) lorsqu'ils sont perçus comme des lieux géographiques etloc.addr.post : les adresses postales.
- **Groupe géo-socio-politique :**un même mot peut parfois être utilisé pour désigner à la fois une région, son peuple et son gouvernement. Ce groupe peut être subdivisé en 3 sous-classes, à savoir : gsp.org : lorsque les régions administratives sont considérées comme une organisation ;gsp.loc: lorsque les régions administratives sont considérées

[9]http://technolangue.net/

[10]http://www.elda.fr/en/

comme un lieu et gsp.pers : lorsque les régions administratives désignent leurs habitants, leur population.

- **Bâtiment et construction humaine :** cette catégorie désigne toutes les entités qui se limitent aux bâtiments et autres constructions fonctionnelles permanentes humaines (en anglais facilities). Il s'agit des lieux renfermés, où l'on peut circuler, à titre d'exemple les maisons, les usines, les stades, les entreprises, les prisons, les musées, etc.
- **Production humaine :** cet ensemble d'entités est répartie en 4 sous-classes, à savoir : prod.award qui comporte tous les types de récompense; prod.vehicule contenant tous les moyens de transport; prod.art désignant toutes les œuvres artistiques etprod.printing qui concerne toutes les œuvres littéraires.
- **Temps :** ce groupe d'entités représente toutes les expressions relatives au temps, divisé en 2 sous-classes : date (date absolue, date relative), désignant tous les mots ou groupes de mots indiquant les jours de la semaine, les mois, les années et les évènements calendaires et heure représentent tous ce qui est en relation avec les horaires.
- **Quantités :** cette catégorie comporte 7 sous-classes de montants y compris âge, durée, température, dimension, poids, vitesse, et valeurs monétaires.
- **Inconnue ou Incertain :** est utilisée pour annoter les entités ne correspondant à aucune des catégories énumérées ci-dessus.

La deuxième campagne ESTER, a démarrée en fin de janvier 2008 et pris fin en avril 2009. Elle a pour finalité de continuer les travaux de recherche démarrés dans la première édition et de lancer de nouveaux axes de recherche. Elle a été organisée grâce aux efforts de la Direction Générale à l'Armement et l'Association Francophone de la Communication Parlée[11], et ELDA (Evaluations and Languageresources Distribution Agency).

Dans cette deuxième édition ESTER2 (Galliano, et al., 2009), le corpus relatif aux entités nommées comprenait 72 heures d'émissions radiophoniques francophones (France-Inter, France Info, RFI, RTM, France Culture, Radio Classique) manuellement transcrites et annotées en entités nommées. Cette édition se caractérisait par la mise en place d'un jeu de 37 types d'entités nommées, dont 7 catégories principales ont été normalisées. Les catégories

[11]http://www.afcp-parole.org/

« Personne », « Organisations », « Lieux », « Temps », « Quantités » et « production humaine » n'ont pas subi de grandes modifications par rapport à la version précédente, en revanche on constate l'apparition d'une nouvelle catégorie sous le nom de « Fonctions », divisée en 5 sous-groupes à savoir : fonc.pol (politique), fonc.mil (militaire), fonc.admi (administrative), fonc.rel (religieuse), fonc.ari (aristocratique), avec la disparition des deux catégories « Bâtiment et construction humaine » et « Groupe géo-socio-politique ». Ceci représente la version définitive du guide d'annotation dans ESTER.

1. 3. 5 Métriques d'évaluation des Entités Nommées

La reconnaissance des entités nomméesest une problématique qui a suscité l'intérêt de plusieurs campagnes et projets d'évaluation. Les mesures traditionnellement utilisées pour mesurer la performance d'une recherche au sein d'une collection de documents ont été effectivement adaptées au concept des entités nommées.

L'évaluation des systèmes permet de mesurer l'écart entre une annotation de référence (faite par expert) et l'annotation obtenues à partir d'un système de reconnaissance d'entités nommées, elle est généralement faite en se basant sur des indicateurs classiques soit en termes de Rappel et Précision et F-mesure (Rijsbergen, 1979; Grishman, et al., 1996)ou en termes de SER (Slot Error Rate)(Makhoul, et al., 1999). Ceux-ci sont des mesures largement utilisées dans les évaluations en TALN(Grouin, et al., 2011), et permettent de calculer le pourcentage des erreurs faites par le système comparé au résultat idéal.

Rappel (mesure de quantité) : est une évaluation de la couverture du système, correspond au nombre d'entités nommées correctement identifiées par le système rapporté au nombre d'entités nommées idéales contenues dans la référence.

$$Rappel = \frac{\text{Nombre d'entités correctement identifiées}}{\text{Nombre d'entités annotées dans la référence}} \quad \textbf{(1)}$$

Précision (mesure de qualité) : est une évaluation du bruit du système, représente le nombre d'entités nommées correctement identifiées par le système rapporté au nombre d'entités nommées ramenées par le système, à savoir les entités correctement et incorrectement étiquetées.

$$Précision = \frac{\text{Nombre d'entités correctement identifiées}}{\text{Nombre d'entités identifiées dans le corpus}} \quad \textbf{(2)}$$

F-mesure (synthèse de R et P) : est une métrique supplémentaire qui vise à faire une synthèse entre les deux mesures précédentes. Il s'agit de la moyenne harmonique pondérée du rappel et de la précision de manière à pénaliser les grandes disparités entre ces dernières.

$$F - Mesure = \frac{(1 + \beta^2) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}} \quad \textbf{(3)}$$

La valeur accordée au coefficient β permet soit d'équilibrer le rappel et la précision (β = 1), soit d'accorder plus d'importance à l'une des deux mesures : le rappel (β > 1) ou la précision (β < 1):

Dans le champ de reconnaissance des entités nommées, deux éléments font l'objet d'une évaluation : d'abord, le repérage de l'entité nommée via ses bornes fixées (frontières de l'entité) ; ensuite, la définition de la classe attribuée à cette entité (type de l'entité).

Techniquement, les métriques qu'on a citées précédemment ont l'inconvénient d'être moins tolérantes aux erreurs. Surtout, quand il s'agit des entités imbriquées, l'absence de l'entité, qui couvre tous les mots contenus dans celle de référence, invalide toute l'hypothèse, même si des entités internes ont été correctement identifiées.

Prenant l'exemple suivant, où plusieurs hypothèses d'annotation sont possibles :

Phrase de Référence : <pers> Leonardo DiCaprio</pers> a remporté l'Oscar.

Hypothèse N°1 : <pers> Leonardo </pers>DiCaprio a remporté l'<pers> Oscar </pers>.

Hypothèse N°2 : <pers> Leonardo DiCaprio</pers> a remporté l'<pers> Oscar </pers>.

À ce titre, le Slot Error Rate ou le taux par types d'erreurs (SER), au vu de la référence (Makhoul, et al., 1999), utilisé dans la campagne d'évaluation ESTER 2, complète les mesures précédentes, et vise à combiner et pondérer différents types d'erreurs. Le principe de base est de fournir un taux d'erreur sur l'ensemble des entités de référence (R) pour lequel on définit les types d'erreur suivants : les erreurs d'insertion (I), de suppression (S), type (T) et d'extension (E), avec une possibilité d'avoir à la fois une erreur de type et d'extension (TE). Ce calcul est donc fait à partir des entités (ou "slots"), tout en déterminant si les entités relevées en hypothèse peuvent correspondre à des entités en référence, que ce soit partiellement (entités avec des erreurs de type ou de bornes) ou totalement (entités correctes).

Nous pouvons détailler les types d'erreur comme suit :

- **Insertion (I)** : représente le nombre d'entités relevées dans l'hypothèse par le système mais qui n'ont aucun mot commun figurant dans celles dans la référence.
- **Suppressions (S)** : ou encore appelée délétion (D), renvoie le nombre d'entités qui figurent dans la référence mais qui ont été complètement manquées par le système.
- **Type (T)** : désignent le nombre d'entités qui ont été relevés par le système mais qui n'ont pas le même type que celles présentes dans la référence.
- **Extension (E)** : il s'agit de deux entités ayant les mêmes types en référence et en hypothèse, mais qui ne commencent ou ne se terminent pas aux mêmes positions, on parle ici de frontières incorrectes.
- **Type et extension (TE)** : représente le nombre d'entités relevées avec des types et frontières incorrectes.

Le SER peut avoir la forme suivante :

$$Slot\ Error\ Rate = \frac{\mathrm{D} + \mathrm{I} + \mathrm{TE} + 0{,}5 \times (\mathrm{T} + \mathrm{E})}{\text{Nombre d'entités identifiées de la référence}} \quad \textbf{(4)}$$

1. 4 Domaines d'applications

En tant que composant autonome, la tâche de reconnaissance des entités nommées devient également un sujet de recherche et de développement fondamental dans de nombreux domaines d'application et du Traitement Automatique des Langues Naturelles (TALN), ainsi que cela est amplement montré dans l'histoire des campagnes d'évaluation tenues. La détermination des expressions correspondant à des entités nommées se présente en effet comme une étape utile pour la mise en place de ces systèmes. Dans la pratique, nous recensons une multitude de domaines qui ont bénéficié d'une intégration de la tâche REN, ci-dessus quelques exemples :

1. 4. 1 Extraction d'informations

Depuis plusieurs années dans la communauté du traitement automatique du langage naturel, la compréhension automatique des textes a fait l'objet de nombreux efforts de recherches s'intéressant à comprendre le sens global d'un document. En revanche, récemment avec l'apparition d'une grande masse de données textuelles, la recherche manuelle d'information

n'est pas délicatement réalisable. L'extraction d'information est donc devenue un enjeu crucial.

L'extraction d'informations est une discipline assez récente qui ne cherche pas à comprendre un texte dans sa totalité, mais consiste, plus techniquement, à extraire à partir d'une collection de textes un ensemble d'informations pertinentes au regard d'une requête bien précise. Ce qui permet de construire une représentation bien structurée d'un document à l'origine non structuré. Cette tâche est relativement difficile car elle exige une part de compréhension et nécessite des connaissances, des ressources lexicales, sémantiques et conceptuelles adaptées aux documents et au domaine à traiter(Nédellec, et al., 2009).

Le champ de recherche d'extraction d'informations couvre toutes les tâches consistant à extraire des informations structurées à partir de textes, à savoir l'extraction d'entités nommées (EN)(Paik, et al., 1996). Plusieurs travaux de recherche entamés dans ce sens attestent l'utilité de cette tâche dans l'EI, permettant d'améliorer la pertinence des résultats de recherche(Guo, et al., 2009).

1. 4. 2 Traduction Automatique

Suite à l'émergence des nouvelles technologies, la taille des données multilingues à traiter devient de plus en plus énorme, ce qui amplifie les besoins en outils de traitement automatique permettant l'édition de données multilingues. A cet effet, plusieurs systèmes ont été mis en place, dont la majorité se base sur l'exploitation des entités nommées(Babych, 2003; Al-Onaizan, et al., 2002; Chen, et al., 2003) et en particulier pour la traduction automatique.

La traduction automatique (TA) désigne, au sens littéral, la traduction des textes d'une langue source (texte originale) vers une langue cible (texte traduit), en se servant d'un ou plusieurs programmes informatiques(Loffler, 1996).

Dans le cas classique, cette tâche repose sur deux processus complémentaires à savoir : la translittération et la traduction des ENs. La translittération consiste en la conversion des signes d'entités nommées d'un système d'écriture à un autre, et ceci indépendamment de la prononciation. En d'autres termes, aucune traduction n'est impliquée dans ce processus et l'opération peut être réversible. Une fois ce travail de translittération est terminé, s'impose la phase la plus importante, à savoir la traduction des entités nommées qui consiste quant à elle à la mise en correspondance des indicateurs des ENs, on parle ici du sens porté sur les ENs.

Grâce à l'intégration d'un système de REN, la traduction automatique peut faire une distinction entre expressions à translittérer et expressions à traduire, elle sert généralement à lever les ambiguïtés relatives à la polysémie. Prenant l'exemple emprunté d'Ehrmann (Ehrmann, 2008):

Texte source : Jack London was an Americanwriter.

Texte cible : Jack Londres était un auteur américain.

La traduction de l'anglais vers le français du mot London, peut être sous plusieurs formes : Londres comme étant la capitale de l'Angleterre (Lieu) et London comme étant un nom de famille de Jack London (Personne). Dans cet exemple, le nom Jack London pour lequel une traduction par Jack Londres est non pertinente. Cette erreur aurait dû être évitée s'il y avait un système de REN permettant d'annoter Jack London comme étant une entité de type personne.

1. 4. 3 La Résolution de Coréférence

La tâche de résolution de coréférence (ou encore nommée le traitement des chaînes anaphoriques), considérée comme une tâche interne du TAL, a été introduite lors de la dernière édition MUC-6 (Grishman, et al., 1996)et dans le cadre des campagnes ACE. Cette tâche s'inscrit d'une façon claire et précise sous le domaine d'extraction d'informations, qui peut être perçues comme un traitement fondamental, nécessaire pour de nombreuses applications du traitement de la langue.

D'un point de vue linguistique, deux mots ou séquences de mots font référence à la même chose. Cela dit que les deux se réfèrent à la même personne, un lieu, une chose ou autres.

Techniquement parlant, la coréférence est une relation linguistique entre deux unités lexicales ou plusieurs unités qui s'appuie sur la tâche de REN grâce à :

- La détermination d'une partie des éléments de la chaîne référentielle, qui peut être sous plusieurs formes à savoir : les noms propres (Abdelilah Benkiran), les groupes nominaux (chef du gouvernement marocain) ou encore les simples pronoms (il),
- La catégorisation des entités nommées : un typage des EN se juge nécessaire et utile à la résolution anaphorique, prenant l'exemple de l'énoncé suivant : « Ahmed s'est levé tôt le matin puis il est parti faire du sport ». Dans cet exemple, l'unité lexicale « Ahmed » et le pronom « il » réfèrent à la même entité, même si les unités lexicales

ne sont pas identiques. Cependant, l'ambigüité sémantique a été relevé grâce au type sémantique de l'entité en question, le pronom « il » devient interprétable.

Les résultats de la tâche de REN interviennent fortement alors dans le processus de résolution de coréférence, les regroupements référentiels pouvant être bruités ou incomplets en fonction de la qualité des mentions fournies.

1. 4. 4 Analyse Syntaxique

L'analyse syntaxique cherche à démontrer la structure hiérarchique des phrases d'un texte et joue un rôle important dans la compréhension du texte. Elle n'a évidemment aucun sens si les données textuelles sont présentées sous forme de métadonnées, à l'inverse des données affichées sous forme de phrases.

La reconnaissance des entités nommées peut constituer un module très utile pour la mise en place d'un analyseur syntaxique robuste comme il a été montré dans les travaux de (Brun, et al., 2004) et (Osenova, et al., 2002).La REN permet d'obtenir ainsi des informations de segmentation et d'annotation au niveau des parties du discours permettant de ne pas gaspiller du temps dans des analyses non pertinentes.

S'agissant d'entités nommées, la virgule et le point, par exemple, ne désignent pas tout le temps des séparateurs, mais peuvent aussi être des parties intégrales d'une entité de type personne (Mrs. Robinson) ou organisation (Snap Inc.).

1. 5 Conclusion

Dans ce chapitre, nous avons présenté une introduction générale de la tâche de reconnaissance d'entités nommées. Nous nous sommes intéressés à l'objet « entité nommée » d'un point de vue théorique, avec pour objectif de proposer une définition de ces unités lexicales. Nous avons, ensuite, fait un tour d'horizon des campagnes et projets d'évaluation menés jusqu'à nos jours. Par le biais de ces derniers, la tâche REN s'est rendue indispensable pour diverses applications du TAL.Un état des lieux des différentes approches de la REN fera l'objet du chapitre suivant.

Chapitre 2 :
Approches de Reconnaissance des Entités Nommées

Dans le présent chapitre, nousprésentons les approches et les systèmes conçus pour répondre aux besoins de la tâche de reconnaissance des entités nommées. Nous allons décrire le principe de base de chaque approche, les algorithmes les plus utilisés et donner des exemples de systèmes existants pour chacune de ces types d'approches. Cette partie va permettre de tracer l'évolution de la tâche de la REN à travers les performances de ces systèmes et de relever les nouveaux défis.

2. 1 Aperçu général sur les approches de reconnaissance des entités nommées

La REN constitue un champ de recherche très actif depuis de nombreuses années dans plusieurs langues. Elle s'est imposée comme un module fortifiant et intégral dans différentes applications du TAL. Ce besoin a encouragé la mise en œuvre de nombreux systèmes de REN permettant de relever les unités lexicales pertinentes dans un texte (Poibeau, 2005). Introduite lors de la série des campagnes d'évaluation MUC (cf.1. 3. 1), plusieurs systèmes ont été proposés basés sur différentes approches.

À l'instar des approches existantes en extraction d'information(Appelt, et al., 1999), les approches utilisées en reconnaissance des entités nommées se distinguent traditionnellement en trois grandes familles(Nadeau, et al., 2007): Les méthodes à base de règles, généralement implémentées sous la forme d'une grammaire formelle construite par la main, et de listes (listes d'entités, dictionnaires, etc.). Ensuite, les méthodes à base d'apprentissage statistique permettant d'apprendre des modèles d'analyse de textes à partir d'un corpus annoté auparavantet d'établir automatiquement une base de connaissances à l'aide de plusieurs modèles numériques comme le CRF (ConditionalRandom Fields), SVM (Support Vector Machine), etc. Au-delà de ces deux grandes approches, il existe une autre nommée hybride qui représente une combinaison entre ses antécédentes.

Néanmoins, quelle que soit l'approche adoptée, il convient de rappeler que le système développé dépend des caractéristiques du corpus en question, notamment quand il s'agit d'un système à base d'apprentissage, étant qu'un modèle généré sur un corpus ne peut pas être appliqué sur un autre, comme le soulignent (Ferrández, et al., 2006).

Dans cette partie, nous donnons un récapitulatif des différentes approches existantes permettant de traiter la problématique de la reconnaissance des entités nommées.

2. 2 Approches à base de règles

L'approche à base de règles, nommée aussi approche linguistique ou approche symbolique, elle est utilisée par la majorité des systèmes de reconnaissance d'entités nommées(Abraham, 2005). Elle s'appuie sur l'intuition humaine (expert-linguiste) dans la construction manuelle des patrons linguistiques, traditionnellement sous forme d'une liste de règles contextuelles (handcraftedrules), qui spécifient les contextes d'apparition de telle entité, ainsi que sur des listes de mots déclencheurs ou ce qu'on appelle les marqueurs lexicaux (trigger words), ou parfois sur un étiquetage syntaxique et finalement sur un ensemble de ressources généralement construites d'un ou plusieurs lexiques (gazetteers) pour l'aide à l'extraction des entités(McDonald, 1993).

En ce qui concerne les marqueurs lexicaux, il s'agit de signes ou d'indices qui entourent l'entité nommée et qui permettent souvent de dévoiler sa présence (ex. Mme. pour Madame ou Inc. pour Incorporated). D'un autre côté les gazetteers (dictionnaires de noms propres) peuvent rassembler classiquement une liste des noms et des prénoms les plus fréquents, des noms de localisations (villes, pays, fleuves, etc.) et parfois des noms d'organisations (organismes, institutions, compagnies, etc.), collectés à partir de ressources externes. Ces gazetteers sont fréquemment utilisées dans les systèmes à base de règles comme elles peuvent servir pour les systèmes à base d'apprentissage.

Cette approche présente de nombreux avantages. Elle est caractérisée par une certaine flexibilité, étant donné que ce sont les experts qui construisent les règles tout en maîtrisant les moyens de les implémenter (expressions régulières, algorithmes informatiques, etc.). En outre, la possibilité de révision et d'actualisation des règles peut se faire d'une manière très simple et pratique, ce qui leur permet de pouvoir tenir compte, à la fois, des cas très fréquents et ceux très rares dans le corpus utilisé pour générer les règles. Ceci permet d'offrir des

résultats de qualité. D'autre part, à l'inverse des approches à base d'apprentissage statistique, elles n'exigent pas forcément la présence d'un corpus annoté(Hewavitharana, et al., 2013).

Parmi les inconvénients majeurs des systèmes de reconnaissance à base d'approches de règles est la mise à jour permanente des ressources linguistiques. Pour avoir un système de REN valide, les différentes ressources employées pour réussir le processus de reconnaissance doivent être constamment actualisées et étendues par de nouvelles entrées (noms et expressions), surtout quand il s'agit de les adapter à de nouveaux domaines, elles sont très liées au domaine ayant servi au développement des règles. En plus, il devient pénible de construire des règles qui couvrent la majorité des formes possibles de présentation des entités nommées.

En termes d'entités nommées, ce type d'approche linguistique a été largement répandu, voire prédominant durant les années 1990, au temps des premières conférences MUC. Pour les langues peu dotées, cette première approche semble la plus convenable.

Les premiers travaux sur la reconnaissance des ENs reposent sur des méthodes à base de règles. Dans le cadre de ces approches, nous citons quelques systèmes populaires à savoir :

Mesfara mis en place un système de REN pour la langue arabe.Ce système est basé sur la combinaison d'analyseur morphologique et d'un système de reconnaissance utilisant des grammaires locales implémentées sous la plateforme linguistique NOOJ(Mesfar, 2008). L'architecture de ce système est schématisée comme suit :

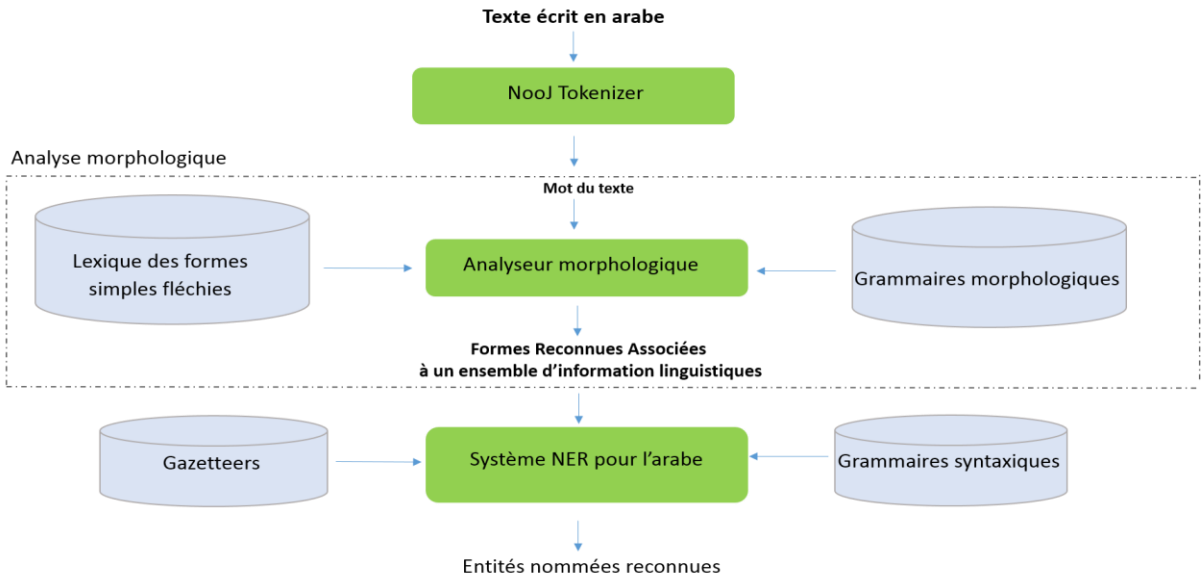

Figure 2-1 Architecture générale du système de REN

Dans la même lignée, Shaalan et Raza ont développé un premier système de REN nommé PERA (Person NarneEntity Recognition for Arabic) (Shaalan, et al., 2007), ensuite NERA (NamedEntity Recognition for Arabic) (Shaalan, et al., 2008) permettant d'extraire dix types d'EN. Ces systèmes se repose sur l'utilisation d'un ensemble de dictionnaires d'EN et sur une grammaire sous forme d'expressions régulières pour la reconnaissance des EN. Le meilleur taux F-mesure acquis par ce système est 98.6%. Dans (Shaalan, et al., 2009), les auteurs ont développé leur système en utilisant une technique de filtrage, qui permet de filtrer les résultats d'un extracteur d'EN à l'aide de métadonnées, par le biais d'une « liste noire » des EN mal formées.

Suivant presque le même principe, Zaghouani et al.,ont présenté un module de repérage des EN à base de règles pour la langue arabe appelé RENAR (Repérage des Entités Nommées Arabes), un outil de repérage des EN à base de règles créées typiquement pour la langue arabe(Zaghouani, 2012). Le diagramme représentant l'architecture du système RENAR est illustré dans la figure suivante :

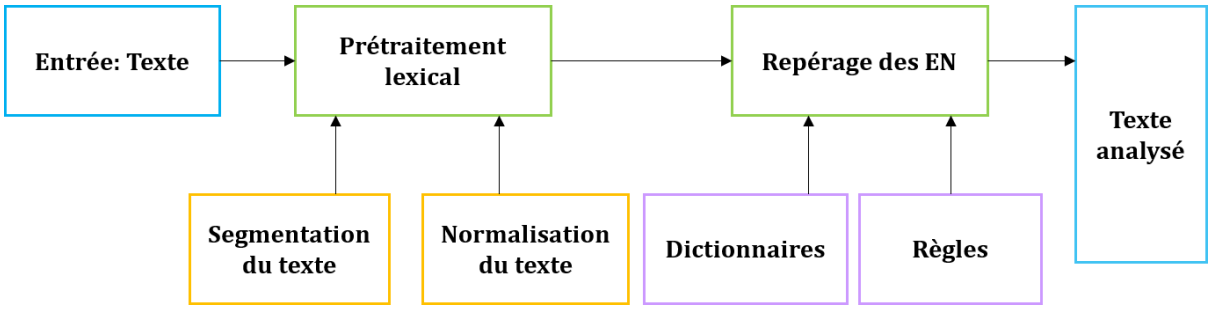

Figure 2-2 Architecture du système RENAR

A partir de cette illustration, on peut remarquer que le processus de repérage commence, d'une part, par une phase de prétraitement lexical qui prépare le texte brut pour son analyse linguistique, en divisant en premier lieu le texte en phrases, ensuite en normalisant son orthographe. D'une autre part, l'opération de détection des EN se fait sur deux étapes. La première repose sur la consultation directe du lexique qui se compose de plusieurs dictionnaires, alors que la deuxième consiste à utiliser des fichiers de règles écrites à la main sous forme d'expressions régulières qui permettent de détecter les EN. Ce système a été évalué sur un corpus de textes journalistiques extraits des journaux arabes mis en ligne. Ainsi, l'outil RENAR a obtenu un Rappel de 59,56%, une Précision de 84,66% et la valeur de F-mesure apportée par ce système est de 69,92%.

Dans le même contexte des langues peu dotées nous évoquons le premier système élaboré pour la langue Iban (Malaisie) par (Fong, et al., 2011). Il consiste à employer des listes des

noms propres et d'un ensemble de règles rédigés manuellement pour permettre l'extraction des entités nommées Iban. Ils ont travaillé sur un corpus collecté à partir de « TunJugahFoundation », ils ont obtenu un F-mesure qui est égal à 76,4%.

2. 3 Approches à base d'apprentissage

2.3. 1 Principe

L'apprentissage automatique est une activité qui a pour objectif de mettre en place des outils informatiques capables de traiter des données juste en se basant sur l'observation d'un ensemble de données en faisant émerger des régularités qui sont ensuite modélisées et les exploiter dans de nouvelles données similaires aux premières. Le processus devient donc fabuleusement réduit, puisqu'il ne nécessite pas une forte maitrise des formalités du domaine en question. L'exploitation de cette activité en TAL a cru exponentiellement depuis plus d'une vingtaine d'années.

Le traitement automatique des données complexes et hétérogènes est un processus crucial pour diverses applications de REN. L'explosion du flux de textes journalistique et l'enrichissement du vocabulaire rend la tâche plus compliquée par les approches linguistiques. À l'opposé de ces derniers qui requièrent souvent plus de maitrise de la langue en question et des connaissances approfondies du domaine pour la construction des règles ou patrons linguistiques, les approches statistiques se caractérisent par leur robustesse et indépendance.

Les approches à base d'apprentissage, ou encore appelée approches statistiques, abordent la reconnaissance des entités comme étant un problème de classification. Il s'agit d'« apprendre », sur de larges corpus de texte où les entités ont été auparavant annotées en assignant chaque entité à la classe à laquelle elle appartient.Des modèles d'analyse numérique pouvant être appliqués sur d'autres corpus. Ces modèles peuvent prendre des aspects différents, d'arbres de décision, modèles probabilistes ou encore chaînes de Markov cachées(Nadeau, et al., 2007).

Ces approches ont été conçues pour avoir une certaine intelligence lors de la prise des décisions. Ce sont principalement certains paramètres qui peuvent être manipulés dans le but d'améliorer les résultats du système, ce qui n'est pas le cas pour les approches symboliques qui n'appliquent que les règles préalablement injectées. Ces approches présentent l'avantage

d'être plus flexibles quant à leur adaptation à une tâche similaire mais portant sur un autre domaine et d'être plus robustes sur des corpus bruités.

Ainsi, on distingue trois principaux types d'approches par apprentissage :

- Les approches supervisées nécessitant un fort degré de supervision ;
- Les approches semi-supervisées nécessitant un degré de supervision moyen à faible ;
- Les approches non supervisées qui ne nécessitent aucune supervision.

Nous allons détailler dans cette section ces différents types d'approches.

2.3. 2 Apprentissage Supervisé

Dans l'approche automatique par apprentissage supervisé (ou discrimination), il s'agit d'une méthode d'apprentissage qui s'appuie sur l'utilisation des exemples prédéterminés sous forme d'un corpus d'entrainement préalablement préparé et annoté, ceci dans le but de pouvoir prendre une décision de classification sur les exemples de test. Chaque exemple doit être décrit par un ensemble d'attributs (en anglais features) qui doivent être suffisamment pertinents afin d'acquérir de bons résultats de classification. A noter que la performance du système augmentera proportionnellement avec la quantité et la qualité du corpus d'apprentissage et dépend aussi de la conception d'un ensemble pertinent d'attributs.

Ce type d'apprentissage se déroule en deux étapes fondamentales : la première consiste à fournir le corpus d'apprentissage annoté, alors que la deuxième repose sur une mise en place des règles permettant de déterminer les classes des entités nommées dans un texte. Ce type requiert une plus grande intervention humaine à chacune des étapes de l'opération d'apprentissage.

L'avantage majeur de ce type, c'est que la présence d'un expert-linguiste n'est pas imposée.Autrement dit, l'annotation d'entités nommées ne requiert pas d'expertise linguistique et de connaissances profondes dans ce domaine mais peut également être effectuée par des non linguistes.

Néanmoins, et comme nous l'avons mentionné au début, afin que les résultats soient bons, le corpus annoté doit être d'une taille importante, ce qui implique un investissement considérable en temps. En outre, similairement aux approches linguistiques, cette méthode est peu portable. Il devient difficile de l'adapter à de nouveaux domaines ou langues, dans ce cas un nouveau corpus de grande taille devra être annoté.

L'approche automatique par apprentissage supervisé se base sur des algorithmes d'apprentissage pour apprendre, à partir d'un corpus annoté, un modèle permettant d'étiqueter les textes. Parmi les algorithmes les plus populaires sont les approches discriminantes, on trouve les approches à base de modélisation graphiques probabiliste comme les chaînes de Markov cachées (en anglais Hidden Markov Models, HMM) (Bikel, et al., 1997), et les champs conditionnels aléatoires (en anglais ConditionalRandom Fields, CRF)(McCallum, et al., 2003), l'entropie maximale (en anglais Maximum Entropy, ME) (Bender, et al., 2003)et machine à vecteurs de support (en anglais Support Vector Machines, SVM) (Takeuchi, et al., 2002). Nous illustrons dans ce qui suit, quelques techniques populaires :

- **Les arbres de décision (en anglais DecisionTrees, DT) :**

Il s'agit d'un des algorithmes les plus populaires de l'apprentissage supervisé. Similairement à un arbre naturel, un arbre de décision se compose de nœuds racines (nommés aussi nœuds internes), desquels partent un ensemble de branches, jusqu'à atteindre des feuilles (nœuds terminaux). Pour arriver aux feuilles, il y a une trajectoire unique partant à partir du nœud racine, à travers les branches. Le parcours dans l'arbre s'arrête dès qu'une feuille est atteinte. Chaque branche dans l'arbre correspond à une décision à prendre, dispose ainsi d'un certain poids et d'une certaine probabilité pour prendre une décision donnée et tous les nœuds terminaux correspondent à une valeur possible de cette requête, représentant des classes.

Les arbres de décision requièrent la création d'un corpus d'entraînement. Ils utilisent un ensemble des variables discriminantes afin de distribuer l'ensemble des mots des exemples dans des groupes homogènes. Les variables discriminantes produisent une série de tests sur la fréquence des mots dans le document. Ces tests permettent au final une classification hiérarchique des mots. Ils concernent les différents descripteurs des mots(Quinlan, 1991).

Un des avantages de ces algorithmes est la possibilité de produire automatiquement un ensemble de décisions sous forme de règles facilement interprétables et lisibles grâce à leur structure arborescente. De plus, elles se caractérisent par leur fluidité et leur simplicité. Elles peuvent être adaptées à d'autres domaines spécifiques. Paliouras et al., ont montré, dans leurs travaux de recherche, que l'emploi de l'arbre de décision pour la génération des règles de décision peut dans certains cas surpasser les systèmes à base de règles(Paliouras, et al., 2000). Néanmoins, cette méthode risque d'être moins performante, quand on a des données bruitées

et des points aberrants qui peuvent se générer pour les classes d'entités rares, aussi en raison de la quantité minime d'un corpus d'entrainement.

Plusieurs algorithmes permettent de représenter les arbres de décision, nous citons à titre d'exemple : ID3 (Inductive DecisionTree) et son successeur C4.5, CART (Classification and RegressionTree), CHAID (Chi-Square Automatic Interaction Detection), QUEST (Quick, Unbiased, Efficient StatisticalTrees). Toutefois, L'algorithme le plus connu et utilisé pour construire des arbres de décision est le C4.5.

- **Les champs aléatoires conditionnels :**

Les CRFs**(**en anglais ConditionalRandom Fields, CRF) font partie des approches à base de modélisation graphiques probabiliste. Ils sont très utilisés notamment dans le domaine du TAL. Ils donnent ainsi de très bons résultats, parfois les meilleurs, dans plusieurs domaines, à titre d'exemples : la tâche de reconnaissance d'entités nommées(McCallum, et al., 2003), l'extraction d'information (Pinto, et al., 2003), l'analyse morphosyntaxique (Sha, et al., 2003) ou l'étiquetage de discours (Altun, et al., 2003).

En effet, le formalisme des champs aléatoires conditionnels (CRF)(Sutton, et al., 2012; Lafferty, et al., 2001) est un modèle graphique linéaire non dirigé, s'inscrivant dans un cadre probabiliste. Ils sont basés sur une approche conditionnelle pour étiqueter et segmenter les séquences de données. Autrement dit, ils cherchent à donner une représentation de la distribution de probabilités d'annotations (étiquettes ou labels) conditionnellement aux observations discrètes à partir d'exemples labellisés (exemples avec les labels souhaités). Sutton et al., ont défini les CRF comme étant une approche qui est « *souvent utilisée pour modéliser les dépendances entre les sorties (probabilités a posteriori d'apparition des classes) d'un premier étage qui fournit des décisions locales (réseaux de neurones, réseaux de neurones récurrents, etc)*(Sutton, et al., 2006)»

L'un des avantages de cette approche, c'est qu'elle permet facilement de retenir le contexte pour étiqueter une séquence de tokens. Bien entendu, elle permet de définir le voisinage, à considérer lors de la prise des décisions, sous une forme graphique.

- **Les réseaux de neurones:**

Les réseaux de neurones (en anglais Neural Network, NN) est une technique qui, de même que pour les arbres de décision, permet, à travers des observations restreintes, de tirer ou d'induire des généralisations plausibles. Il constitue une mémoire lors de sa phase

d'apprentissage à partir d'un jeu d'exemples déjà annotés. En matière d'informatique, on appelle réseau de neurones un ensemble d'entités (ce qu'on appelle les neurones) complètement interconnectées.

Un réseau de neurones comporte (cf. Figure 2-3) trois types de couches :

- La couche d'entrée : représente le vecteur d'observations, auquel on attribue une activation en fonction des données que le réseau doit traiter.
- La/les couche(s) cachée(s) : contient l'ensemble des neurones et leurs poids associés, à travers laquelle l'activation des neurones d'entrées se reproduit et se modifie.
- La couche de sortie : désigne l'ensemble des classes possibles du problème posé en entrée.

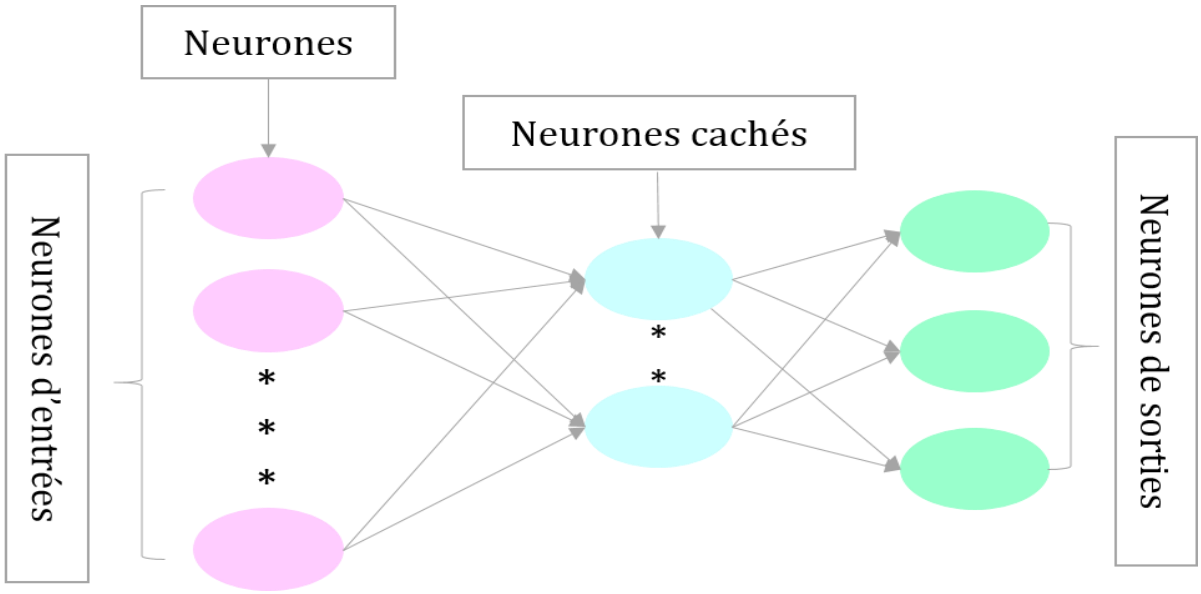

Figure 2-3 Schéma général d'un réseau de neurones

A noter que la phase d'activation (état d'excitation) dépend des neurones situés en amont, ainsi que de la force des liens qui les relient.

Ils représentent une structure formée de suite successive de couches de nœuds et qui permet de déterminer une fonction de transformation non linéaire des vecteurs d'entrées en vecteur de sorties sous forme de classes possibles. Pratiquement, l'apprentissage consiste à faire passer les données annotées en entrée du réseau et à modifier les petites erreurs qui se produisent tout en modifiant les nœuds internes du réseau, par la suite au fur et à mesure le processus finit par classer convenablement les entrées en question.

La présence des neurones dans le réseau de même que le nombre de couches utilisées influent considérablement sur le résultat de classification. En outre, les réseaux de neurones ont l'inconvénient d'être coûteuse en termes de temps en phase d'apprentissage.

A l'exemple de la tâche de la REN, l'entrée du réseau contient l'EN à classer. La couche de sortie, de son côté, correspond à l'ensemble des classes. Après excitation du réseau, les valeurs de la couche de sortie désignent les classes possibles du document.

Nous allons présenter en détaildans le chapitre 4un autre algorithme plus connu dans cette catégorie qui est le SVM.

2.3. 3 Apprentissage non Supervisé

Il existe aussi des méthodes par apprentissage automatique de type non supervisé ou encore dites « clustering », qui, selon l'histoire, ont fait de rapides progrès au cours des dernières années et permettent aussi d'avoir des résultats honorables. Elles consistent à apprendre à classer sans supervision et automatiquement, qui comme leur nom l'indique, ils sont l'opposé des précédentes. Ils n'ont pas besoin d'avoir un corpus labellisé, les règles apprises sont évaluées grâce à des heuristiques(Collins, et al., 1999).

Selon Candillier(Candillier, 2006) l'apprentissage non supervisé « *consiste à former différents groupes à partir d'un ensemble de données, de telle manière que les données considérées comme les plus similaires soient associées au même groupe et qu'au contraire les données considérées comme différentes se retrouvent dans des groupes distincts, permettant ainsi d'extraire de la connaissance à partir de ces données* ».

Dans ce cas, l'apprentissage se ramène alors à cibler les regroupements homogènes (ou clusters) d'éléments existants dans le corpus, qui se ressemblent et qui sont séparés des autres éléments différents situés dans d'autres regroupements. Les ENs sont groupées automatiquement en sous-ensembles selon la similarité de leur contexte immédiat. La similarité est généralement calculée selon une fonction de distance entre paires d'exemples. Il s'agit ici d'une forte similarité intra-classe, et une faible similarité interclasses(Kaufman, et al., 2005).

Au début de l'opération d'apprentissage, nous n'avons aucune information ni sur la description des classes sémantiques, ni sur leurs nombres. C'est l'algorithme de clustering qui va déduire toutes ces informations. Nous n'avons pas non plus des données en entrée qui sont déjà labellisées, typiquement, c'est l'algorithme qui se charge de déterminer par lui-même la

structure de ces données et de construire des clusters dont les caractéristiques sont les mêmes. Il est également très intéressant d'utiliser ce type d'apprentissage lorsque nous ne savons pas ce que nous cherchons.

On peut prendre en compte que chaque EN ne peut faire partie qu'à un seul cluster, ou encore elle peut avoir une probabilité d'appartenir à d'autres clusters créés. Il s'agit dans le premier cas d'un « hard-clustering » et dans le deuxième de « soft-clustering ».

L'efficacité des méthodes de clustering dépend de ces trois éléments :

- La manière de calculer la similarité entre les éléments à regrouper.
- La manière de construire les clusters.
- La manière de définir le nombre de clusters à construire.

Dans la littérature, il existe plusieurs types d'algorithmes d'apprentissage non supervisé tels que les algorithmes reposant sur des arborescences (partitionnement, partitional clustering) et les algorithmes de classification hiérarchique (clustering hiérarchique, hierarchical clustering). Le plus populaires des algorithmes d'apprentissage non supervisé est le K-means(Hartigan, 1975).

- **K-moyennes**

K-moyennes (en anglais k-means) est un algorithme de partitionnement qui permet de partitionner un ensemble de données automatiquement en K nombre de groupes (clusters). Typiquement, ce partitionnement consiste à un regroupement des données tout en se basant sur leur degré de similarités. Il consiste à priori de choisir k points qui désignent les centres des clusters à créer, ensuite d'assigner les autres points aux centres les plus proches. Cette assignation est faite en se basant sur le calcul de distance entre les points. Ensuite, une phase de raffinement des clusters s'exige, cette phase repose sur un recalcul, d'une manière itérative, des centres de cluster et sur une réassignation des points de clusters. Une fois que l'algorithme constate une stagnation au niveau des points, il s'arrête.

Une liste exhaustive des inventaires sur les méthodes non supervisées sont proposés dans (Berkhin, et al., 2002; Xu, et al., 2005).

2.3. 4 Apprentissage semi-Supervisé

Quand on ne dispose pas de suffisamment d'exemples étiquetés, la performance des algorithmes d'apprentissage supervisé s'atténuent considérablement. L'apprentissage semi-

supervisé, parfois appelé légèrement supervisé (en anglais weaklysupervised) a fait son apparition pour remédier à ce problème.

Il s'agit d'un type d'apprentissage qui se distingue de l'apprentissage non supervisé par le fait que l'algorithme procède à un apprentissage avec une intervention humaine moyenne à minimale et ne nécessite qu'un nombre limité de données injectées préalablement afin de fonctionner correctement.

En termes d'entités nommées, son principe est d'utiliser un petit ensemble d'exemples d'entités étiquetés (quelques noms de personnes et d'organisation ou de lieux) afin de les injecter dans le système qui procède alors à l'analyse des phrases contenant ces types d'entités nommées dans des exemples non étiquetés. Par la suite, le système va pouvoir extraire les caractéristiques contextuelles pour chaque type d'entités. Après plusieurs itérations, un grand nombre d'entités nommées, qui apparaissent dans des contextes similaires, peut être localisé. Cette opération itérative permet à chaque itération de raffiner la description des contextes et de ne prendre en considération que les indices discriminatifs. Le résultat de cette analyse peut être appliqué sur un grand ensemble d'exemples non annotés.

Les techniques les plus connues dans cette catégorie sont les Bootstrap (en anglais bootstrapping), qui consiste à s'amorcer à partir d'un petit jeu d'exemples et de l'élargir par itérations successives en exploitant divers descripteurs comme les relations syntaxiques (Cucchiarelli, et al., 2001)ou synonymiques (Paşca, et al., 2006).

Les systèmes à base d'apprentissage se sont largement amplifiés ces dernières années, ont acquis une attention grâce à leur facilité de mise en œuvre.

Dans une perspective de reconnaitre les ENs pour la langue arabe en se servant d'un ensemble de particularités de cette langue, une série de travaux basée sur une approche à base d'apprentissage a été entamée par Benajiba et ses co-auteurs (Benajiba, et al., 2007). Ils ont travaillé, notamment et en premier lieu, sur la construction du corpus ANER (ou ANERcorp) qui contient plus que 150 000 tokens annotés spécialement pour cette tâche, ensuite les gazetteersANERgazet pour développer le système ANERsys.

Dans une première tentative, (Benajiba, et al., 2007) explorent un étiquetage morphosyntaxique fondé sur l'algorithme d'apprentissage statistique de maximum d'entropie, dans le but de pouvoir reconnaitre les noms propres longs. Il est à noter que le corpus d'apprentissage du système ANERsys est de 125 000 mots et le corpus de test est de 25 000 mots. Le système a donné des résultats assez encourageants avec une F-Mesure égale à

55.23%. Toutefois, le problème majeur réside dans le cas où une EN est composée de plusieurs mots. Cette approche est étendue ensuite en décomposant la prédiction en deux sous-tâches : d'abord les frontières de l'EN en introduisant des catégories morphosyntaxiques (POS), puis à la détermination de son type.

Une seconde approche, fondée sur la combinaison des CRF et SVM (Benajiba, et al., 2008).Le système a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, intègre les caractéristiques lexicales, syntaxiques et morphologiques, ce qui a amené à de meilleures performances, fondamentalement en termes de rappel.Benajiba et al., prouvent également l'efficacité d'un prétraitement des textes pour séparer les différents constituants du mot (proclitiques, lemme, et enclitiques)(Benajiba, et al., 2008). Dans sa thèse, (Benajiba, 2009) a essayé de combiner plusieurs types de modèles d'apprentissage comme le SVM, maximum d'entropie et le CRF. Il a pu obtenir de très bonnes performances en combinant plusieurs modèles à la fois, plutôt qu'en utilisant un seul modèle d'apprentissage.

Parmi les travaux de reconnaissance des entités nommées dans divers langues peu dotées, on trouve le système de (Srikanth, et al., 2008) pour le Télugu qui ont utilisé la technique CRF pour extraire les entités nommées et ils ont obtenu un score de f-mesure qui est égal à 92%, (Ekbal, et al., 2008) pour le Bengali. Ils ont réussi à avoir un résultat dont la valeur de f-mesure est de 91,8%, tout en utilisant le SVM.

Similairement aux autres langues précédentes, les travaux de Vijayakrishna(Vijayakrishna, et al., 2008)ont été consacré à la mise en place d'un système de reconnaissance des entités nommées pour le Tamil en se servant du CRF et ils ont obtenu un résultat satisfaisant qui est égal à 80,44%, et il en existe bien d'autres systèmes très pertinents pour la tâche de REN.

2. 4 Approches Hybrides

Au-delà des deux approches citées ci-dessus, une troisième approche a fait son apparition, qualifiée de mixte ou hybride. Elle représente une combinaison entre ses antécédentes, visant à exploiter les avantages et bienfaits de ces dernières, tout en éliminant certains de leurs inconvénients (problème de réutilisation des règles au niveau des méthodes à base de règles, performances réduites pour les approches statistiques avec des corpus d'entraînement de taille réduite, etc.).

Cette dernière approche est sans doute la plus prometteuse, elle utilise des règles écrites manuellement mais construisent aussi une partie de ses règles en se basant sur des informations syntaxiques et des informations sur les discours extraits de données d'apprentissage grâce à des algorithmes d'apprentissage, des arbres de décisions, etc.

Trois principales options sont possibles pour dresser cette approche :

- La première option consiste à générer un modèle de règles en appliquant un apprentissage automatique sur un corpus de textes annotées et par la suite un expert-linguiste intervient pour valider les entités nommées candidates qui ont été relevées.
- Contrairement à la première option, cette deuxième exige, a priori, un ensemble de règles linguistiques construites par un expert-linguiste, puis lors du processus d'apprentissage sur un large corpus de texte, ces règles vont être étendues.
- Quant à la troisième option, elle repose sur les élections entre les deux types de règles : celles construites manuellement et celles engendrées automatiquement à l'aide du processus d'apprentissage sur un corpus annotés. Puis, en se servant des résultats d'apprentissage, le système choisit laquelle des deux prédictions peut retenir.

Parmi les systèmes hybrides les plus cités, on trouve le système Nemesis (Fourour, 2002) qui consiste à construire un système permettant la délimitation et la catégorisation des ENs et plus précisément les noms propres pour le français. Il se base essentiellement sur les indices internes et externes définis par (McDonald, 1996), sous forme de règles de grammaire qui exploitent des lexiques spécialisés combinés avec un système par apprentissage. L'architecture de Nemesis (cf. Figure) comprend principalement quatre modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles, ensuite apprentissage et reconnaissance.

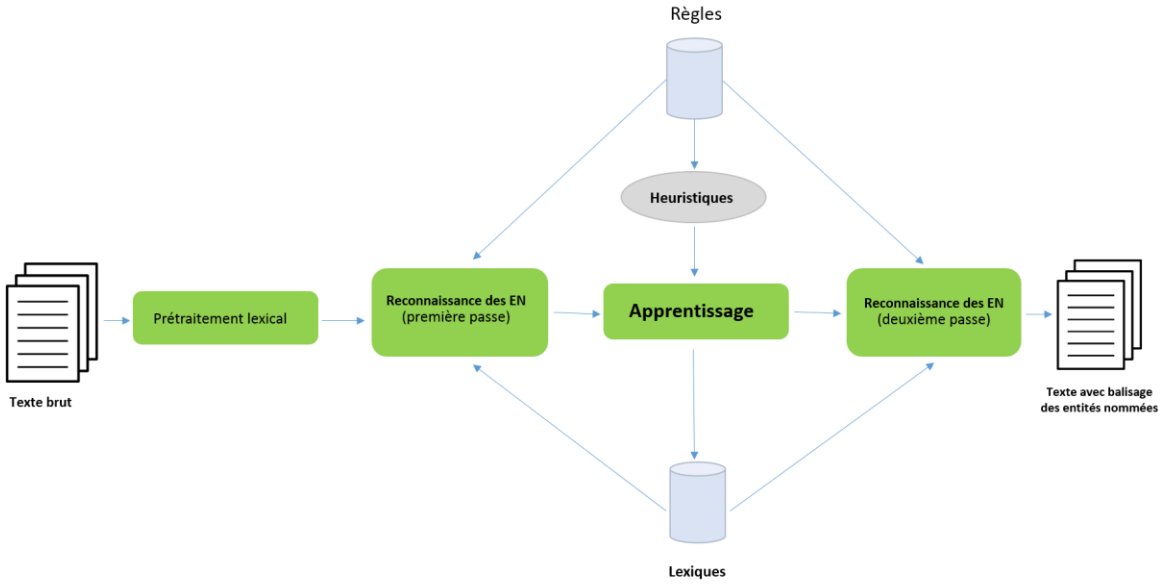

Figure 2-4 Architecture générale du système Nemesis

Selon la Figure 2-4, le système d'apprentissage est employé entre deux processus de reconnaissance par règles, il a pour finalité d'induire de nouvelles règles qui vont former la deuxième passe.

Dans ce système, Fourour a relevé des conflits de chevauchement, d'inclusion et d'accolement (c'est le cas où deux EN sont identifiées, l'une se situe immédiatement à la suite de l'autre) qui sont résolus par la suite soit par l'ajout de nouvelles règles, soit par le changement des priorités des règles ou par la fusion de deux règles.

L'évaluation de Nemesis a été réalisée sur un corpus composé de textes issus du journal « Le Monde et du Web ». Les performances atteintes sur l'ensemble des entités nommées, sont de 95 % pour la précision et 90 % pour le rappel.

Également, dans ce cadre de méthodologie mixte, nous citons le système LTG (Mikheev, et al., 1998) pour les ENs anglaises qui a obtenu la meilleure performance lors de la MUC-7 avec une F-mesure de 93,39 %. L'extraction des entités Enamex par LTG est faite suivant plusieurs étapes à savoir :

- Un passage des règles les plus sûres (sure-firerules) qui produit une première reconnaissance partielle (partial match 1),
- Un passage des règles plus lâches (rule relaxation), qui produit une deuxième reconnaissance partielle (partial match 2), traitement des titres des articles (titleassignment).

Sa particularité est qu'il utilise deux passes de reconnaissance dont la première est un élément d'information pour la deuxième. La première passe est basée sur des règles et une liste de mots déclencheurs pour détecter les entités certaines. Les entités détectées sont ensuite utilisées pour générer des variantes d'entités en changeant l'ordre des mots ou en les supprimant. Cette nouvelle liste combinée avec un algorithme probabiliste fondé sur le modèle d'entropie maximale finalise l'étiquetage.

2. 5 Conclusion

Dans ce chapitre, nous avons pu recenser les diverses approches possibles, qui sont historiquement divisées en trois grandes catégories : approches à base de règles, approches à base d'apprentissage et hybrides. Nous avons ensuite présenté aussi brièvement quelques systèmes existants pour chaque approche. Cet aperçu nous a permis de comparer les différents outils pour la mise en œuvre de ces approches ainsi que les différents jeux de données conçus pour cette tâche.

En décrivant ce chapitre, nous avons constaté que les approches symboliques, étant anciennes, sont les premières à traiter cette tâche. Elles sont considérées comme simples, flexibles et plus précises. Toutefois, elles exigent une connaissance approfondie de la langue des ENs afin de pouvoir mieux mettre au point les règles linguistiques. Suivies massivement par les approches statistiques qui s'avèrent utiles dans le cas où l'utilisation des règles manuelles n'arrive pas à bien reconnaitre les ENs, ces approches nécessitent des corpus d'apprentissage annotés, qui ne sont pas toujours disponibles ou qui doivent être entièrement construits. Parallèlement à cela, nous avons constaté que les approches hybrides représentent un certain compris entre les approches à base de règles et celle à bases d'apprentissage, non seulement en termes de précision et rappel de façon générale mais également en termes de types d'entités ciblées et du domaine d'application, etc.

Bien que notre intérêt s'oriente en particulier vers la question de la reconnaissance des EN notamment dans la langue amazighe, la majorité des systèmes que nous avons illustré ont été dédiés et optimisés pour supporter des langues autres que l'amazighe. De ce fait, Il nous parait fonctionnel et utile de mettre en place les premiers systèmes de REN en se basant sur les différentes approches existantes, tout en tirant partie des divers systèmes conçus pour d'autres langues. Le chapitre suivant est dédié à la présentation d'un bref état des lieux de la

langue amazighe ainsi qu'une description de ses caractéristiques et les travaux qui ont été réalisés jusqu'à présent afin de promouvoir l'informatisation de cette langue.

Chapitre 3 :
Présentation de la langue amazighe marocaine

Après avoir introduit la tâche de la reconnaissance des entités nommées ainsi que ses approches, nous allons dans ce chapitre donner une rapide description de la langue amazighe, son statut géographique, et ses diverses variantes dialectales. Nous allons ensuite présenter ses caractéristiques ainsi qu'un aperçu de sa morphologie. Puis, nous allons recenser les ressources linguistiques qui ont été conçu pour la langue amazighe ainsi que les outils aidant à son traitement, permettant son intégration aux NTIC. Et finalement, nous allons lister quelques éléments entravant l'informatisation et surtout la REN en langue amazighe.

3. 1 Historique de la langue amazighe

La langue amazighe (tamazight, □□□□□□□□), couvre une aire géographique immense, elle est essentiellement parlée en Afrique du Nord, Sahara-Sahel. Elle est également parlée par d'autres communautés dans certaines régions du Niger et du Mali, ainsi que des milliers d'immigrants amazighes partout dans le monde. On peut la considérer comme la langue autochtone du Nord de l'Afrique. La langue amazighe est l'une des branches de la grande famille linguistique chamito-sémitique, ou encore appelée afro-asiatique (Ouakrim, 1995), constituée de plus de trois cents langages, dont certains se sont éteints et d'autres ont été utilisés de façon secondaire comme langue liturgique. Cette famille comprend, outre l'amazighe, le sémitique, le couchitique, l'égyptien (ancien) et, avec un degré de parenté plus éloigné, le groupe "tchadique" (haoussa)(Greenberg, 1963).

Elle a toujours disposé d'un statut restreint et minoré, elle est diversifiée en de nombreuses variétés dialectales dont l'importance démographique va de quelques centaines à plusieurs millions d'individus. Les amazighophones sont présents principalement au Maroc et en

Algérie, au Niger, au Mali et au Burkina-Faso (touareg), en Libye, en Tunisie et aux extrémités du domaine berbère, en Mauritanie et en Égypte.

On peut classifier la langue amazighe en trois grandes zones dialectales capitales dont les trois principales langues vernaculaires peuvent être identifiées :Tarifit (Nord du Maroc), Tamazight (centrale) et Tachelhit (Sud du Maroc). Sur le plan linguistique, on appelle ceci la prolifération des dialectes en raison de facteurs historiques, géographiques et sociolinguistiques.

L'amazighe est longtemps resté sans aucune reconnaissance institutionnelle au Maroc et en Algérie. Le statut de l'amazighe a connu cependant de sensibles améliorations depuis quelques années dans ces deux pays.

Certains États du Maghreb ont créé des institutions spécialisées, telles que l'Institut Royal de la Culture Amazighe (IRCAM)[12] au Maroc et le Haut-Commissariat de l'Amazighité (HCA)[13] en Algérie, dans le but était de standardiser, redonner aux culture et langue amazighes la place qu'elles méritent, ainsi que pour uniformiser les structures et à adoucir les pluralités qui représentent des difficultés au niveau de l'intercompréhension.

Au Maroc, suite au discours royal en 2001, l'amazighe est devenue une langue institutionnelle. Et grâce à la constitution de 2011, l'amazighe a jouit auprès de sa consœur l'arabe d'un statut d'une langue officielle. Tandis qu'en Algérie, l'amazighe est depuis 2002 « langue nationale ». En revanche et depuis quelques années, plusieurs publications dans divers domaines ont connu le jour.

La langue amazighe a été introduite dans le système éducatif comme matière inévitable dans les écoles primaires, un peu plus de 3 000 écoles et plus de 600 000 élèves suivent cet enseignement, quel que soit la langue maternelle des élèves, en perspective d'une généralisation graduelle aux niveaux scolaires, ainsi qu'elle a été intégré dans l'administration et les médias, à savoir une chaîne de télévision amazighe qui a été lancée le premier mars 2010, et au niveau de l'enseignement supérieur, des filières d'études amazighes et des masters ont été créés. La quantité et la qualité de thèses et de mémoires rentrant dans le cadre de la linguistique et la littérature amazighes par rapport à l'ensemble des travaux

[12]http://www.ircam.ma/

[13]hca-dz.org/

universitaires entamé au Maroc, démontre la nouveauté de l'amazighologie, une culture et une langue dignes de considération.

L'amazigheest maintenant une langue qui possède tous ses attributs : dotée d'une graphie officielle, un codage propre dans le standard Unicode, une grammaire, une orthographe ainsi qu'un vocabulaire très riche et une littérature orale fabuleusement fortuné.

3. 2 Caractéristiques de la langue amazighe

3. 2. 1 Ecriture Tifinaghe : Historique

Bien que l'amazighe soit une langue essentiellement à tradition orale, elle fait partie des languesquipossèdent une écriture autonome et indépendante. Depuis au moins deux millénaires et demi, les amazighophones disposent de leur propre système d'écriture appelé « libyco-berbère » ou encore appelé « Tifinaghe».

Comme la langue, l'écriture amazighe n'est pas fondamentalement normalisée, à travers le temps, elle subit un certain nombre de modifications de variations inévitables, passant du libyque jusqu'aux néotifinaghe, en passant par le Tifinaghe saharien et le TifinagheTouareg. Ces variations s'expliquent à la fois par une adaptation aux particularités phonétiques locales et par la durée d'existence de cette écriture qui a induit d'inévitables évolutions et adaptations.

On distingue traditionnellement entre plusieurs transcriptions amazighes, à savoir :

- **Le libyque :** Il s'agit des variétés de Tifinaghe les plus anciennes. Il existe deux formes du libyque, l'oriental et l'occidental. La forme occidentale de son usage se place dans tout le long de la côte méditerranéenne de la Kabylie jusqu'au Maroc et sans doute aux Îles Canaries. La forme orientale couvre le nord de la Tripolitaine, la Tunisie et l'Algérie orientale. La forme occidentale serait plus primitive et compte un plus grand nombre de signes supplémentaires que l'oriental, mais présente aussi plus de variations. En revanche, la forme orientale étant influencée par l'écriture punique.
- **Le Tifinaghe saharien :** Il est également nommé libyco-berbère ou touareg ancien. Il contient des signes supplémentaires par rapport à la transcription libyque, à savoir un trait vertical pour noter la voyelle finale /a/. Cette variété fut utilisée pour transcrire le touareg ancien mais les modalités du passage entre le libyque et le Tifinaghe saharien restent inconnues.

- **Le Tifinaghe touareg** : Il existe au sein du Tifinagh touareg quelques divergences dans la valeur attribuée aux signes qui correspondent aux variations dialectales touarègues. Si d'une région à une autre, la forme et le nombre des signes peuvent changer, les textes restent en général mutuellement compréhensibles. La nouveauté la plus remarquable est la ligature à dernière consonne /t/ ou à première consonne /n/ ; Comme pour le saharien, le Tifinaghe touarègue dispose d'un signe /./ pour noter les voyelles finales appelées Tighratin. Le Tifinaghe touarègueest souvent utilisé pour des transcriptions sur des objets (bijoux, armes, tapis, etc.), pour des déclarations amoureuses et pour des épitaphes. Toute transcription commence par la formule awanäk (c'est moi) + nom + innân (qui a dit).
- **Le néotifinaghe :** Le néotifinaghe représente un système alphabétique standard plus adéquat et utilisable pour tous les parlers amazighes actuels. La revitalisation de l'alphabet amazighe est due au travail incontestable fait par l'AB (Académie Berbère - AgrawImazighen) à la fin des années 60, une association formée par des jeunes militants amazighs (kabyles en grande partie). Plusieurs variantes du néotifinaghe ont été développées à partir des inscriptions anciennes et de l'écriture tifinaghe touarègue. Ce système d'écriture est largement diffusé au Maroc et en Algérie surtout en Kabylie.

L'IRCAM a préconisé le choix du néotifinaghe au lieu des alphabets arabes et latins.

3. 2. 2 Tifinaghe : L'alphabet amazighe

3. 2. 2. 1 Alphabet

L'alphabet amazighe standardisé est basé sur un système graphique à tendance phonologique. Cependant, il ne retient pas toutes les réalisations phonétiques produites, mais uniquement celles qui sont fonctionnelles. Il est composé de 33 caractères, comprend 27 consonnes, 2 semi-consonnes, 3 voyelles pleines et une voyelle neutre :

□, □, □, □□, □, □, □, □, □, □□, □

3. 2. 2. 2 Ponctuation

La ponctuation est essentielle pour structurer la phrase et la rendre intelligible. Dans la langue amazighe, il n'y a pas de signe de ponctuation particulier. L'IRCAM a adopté l'utilisation des signes conventionnels présents dans les écritures latines. On recense traditionnellement

treize signes de ponctuation qui s'insèrent dans le texte : le point {.}, le point d'interrogation {?}, le point d'exclamation {!}, la virgule {,}, le point-virgule {;}, lesdeux-points {:}, les points de suspension {…}, les parenthèses {()}, les crochets {[]}, les guillemets {« »}, le tiret {–}. Grevisse y rajoute la barre oblique {/}. Les accolades { {} } sont également largement utilisées.

3. 2. 2. 3 Chiffres

En ce qui concerne les chiffres, l'IRCAM fait usage de tous les chiffres simples ou composés (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, etc.) et de tous les signes logiques conventionnels (i. e. +, -, =, x, ÷, %, α, β, γ, Σ, π, etc.) pour l'écriture tifinaghe. Cette transcription ne présente donc aucune nouveauté.

3. 2. 2. 4 Ecriture de gauche à droite

Historiquement, l'amazighe des anciennes inscriptions s'écrivait horizontalement de gauche à droite ou de droite à gauche, ou bien verticalement de bas en haut ou de haut en bas. Comme C'est le cas de la majorité des systèmes d'écriture, en particulier les écritures latines et grecques, l'IRCAM a retenu la direction horizontale de gauche à droite pour l'écriture Tifinaghe.

3. 2. 2. 5 Codage de l'amazighe

Afin de conquérir le marché informatique assez important dans le monde, l'établissement d'un code amazighe unifié était une affaire urgente. Certains chercheurs ont cherché, en premier lieu, à définir des jeux codés pour l'alphabet amazighe. Le codage du caractère Tifinaghe donc constitue le premier pas vers l'exploitation des TICs, permettant l'écriture de l'amazighe à travers une représentation numérique pour chaque caractère. Pour ce faire, la première étape a consisté en une adaptation de la norme ISO 8859-1 pour coder les caractères Tifinaghe en codage ANSI, afin de faciliter l'intégration de l'amazighe dans le système éducatif. Quoique, la portée de ce codage privé de l'IRCAM a été limitée, ne permettait pas de représenter l'alphabet amazighe tout complet et la gestion des textes comportant plusieurs systèmes d'écriture était difficile. D'ailleurs, les traitements des textes multilingues doivent jongler à la fois avec les différentes normes de codage et avec les polices associées. D'où la nécessité et l'importance d'intégrer le Tifinaghe dans le plan multilingue, c'est-à-dire un codage plus portable et facile à gérer tel que l'Unicode qui a permis d'affecter un code unique à chaque caractère.

3. 2. 2. 6 Clavier

L'intégration de l'amazighe dans la norme internationale de la prescription des claviers ISO/CEI 9995 a fixé définitivement les claviers de la langue amazighe conçus pour la bureautique. En fait, cette norme a spécifié deux types de claviers : un clavier tifinaghe de base, permettant la saisie des 33 lettres utilisées aux écoles du royaume du Maroc et un clavier tifinaghe étendu, permettant la saisie des 55 lettres tifinaghes.

3. 2. 2. 7 Police de caractère

Afin de faciliter l'intégration de la langue amazighe dans le système éducatif marocain et d'encourager la publication assistée par ordinateur, huit polices de caractères, associées au codage ANSI, ont été proposées. Cette proposition a été suivie par l'élaboration d'une nouvelle génération de polices, cette génération inclut les polices associées à l'Unicode.

3. 2. 3 Morphologie Amazighe

Etant donné que l'amazighe est une langue à tradition orale plus qu'une langue écrite, mettre au point la grammaire de cette langue n'était pas une tâche facile. Il ne s'agit pas de la grammaire particulière d'une variété dialectale ou parler, mais plutôt une grammaire de l'amazighe marocain standard. La langue amazighe présente une morphologie riche. Elle peut être considérée comme une langue complexe. Dans cette partie on présente les propriétés morphologiques des catégories syntaxiques majeures de l'amazighe, en l'occurrence, le nom, le verbe, le pronom et les particules(Boukhris, et al., 2008).

3. 2. 3. 1 Nom

En Amazighe, le nom est une unité lexicale formée d'une racine et d'un schème (modèles de formation des mots permettant d'obtenir des mots à partir de racines abstraites, représentant des notions sémantiques générales ou des significations précises).

Il est à noter que le nom varie en genre (féminin, masculin), en nombre (singulier, pluriel) et en état (libre, annexion). En outre, il peut prendre différentes formes à savoir : une forme simple (□□□□□ [argaz] "homme"), forme composée (□□□□□□□ [buhyyuf] "la famine") ou bien une forme dérivée (□□□□□□□ [amsawaḍ] "la communication").

- **Le genre :** Le nom amazighe connaît deux genres, le masculin et le féminin.

Le nom masculin : il commence traditionnellement par une des voyelles initiales :□ [a], □ [i] ou bien u [u], à titre d'exemple :□□□□ [afus] "main", □□□□□ [argaz] "homme". En

revanche, il existe certains noms qui font l'exception à cette règle, comme par exemple :□□□□ [illi] "(ma) fille" ou □□□□□ [badad] "amour".

Le nom féminin : il prend généralement la forme suivante :□....ⵜ [t....t], celle-ci permet, dans la plupart des cas, d'obtenir le féminin à partir du radical d'un nom masculin, à titre d'exemple : □□□□□ [agmar] "cheval" → □□□□□□□ [tagmart] "jument". Pour certains noms, le féminin est marqué par un autre mot différent, comme : □□□□□ [argaz]"homme" → □□□□□□□ [tamttut]"femme". Il est à noter qu'il y a des noms masculins qui n'ont pas de correspondant féminin comme : □□□□ [adfl] "neige".

- **Le nombre :** Le nom amazighe, qu'il soit masculin ou féminin, possède un singulier et un pluriel. Ce dernier peut prendre plusieurs formes(Oulhaj, 2000): le pluriel externe, interne, mixte et le pluriel en □□ [id].

Le pluriel externe : Le nom ne subit aucune modification interne, il est obtenu par une alternance vocalique ; seule la voyelle initiale a- se transforme en i- ; accompagné par une suffixation de l'indice n [n] ou l'une de ses variantes (□□ [in], □□ [an], □□□ [ayn], □□ [wn], □□□ [awn], □□□ [wan], □□□ [win], □□ [tn], □□□ [yin]), tout dépend si le nom est masculin ou féminin. □□□□□□[asafar] → □□□□□□□ [isafarn] "médicaments, remèdes", □□□□□□ [tabrat] → □□□□□□□□ [tibratin] "lettres".

Le pluriel interne (ou brisé): est obtenue par une alternance vocalique initiale suivi d'un changement de voyelle internes (souvent d'une voyelle) sans aucun ajout de suffixes : □□□□□□ [agadir] → □□□□□□ [igadur] "murs, forteresses".

Le pluriel mixte (suffixation + alternance interne): est formé par une alternance d'une voyelle interne et/ou d'une consonne plus une suffixation par n [n] : □□□□□ [izikr] → □□□□□□□ [izakarn] "cordes" ; ou bien par une alternance vocalique initiale accompagné d'un changement vocalique final a [a] plus une alternance interne (□□□□□□□ [amggaru] ->□□□□□□□ [imggura] "derniers").

Le pluriel en □□ [id] : ce dernier type est obtenu par une préfixation de □□ [id] du nom au singulier. Ce procédé est appliqué à un ensemble de cas de noms à savoir : des noms à initiale consonantique, des noms propres, des noms de parenté, des noms composés en bu - / mmu - ou bab / lal, des numéraux, ainsi que pour les noms empruntés non intégrés (□□□□□□ [lkamyu]"camion" → □□ □□□□□□ [id lkamyu]). Il est a mentionné que certains noms n'ont pas de singulier correspondant comme par exemple : □□□□□□ [idammn] "sang" ou □□□□ [aman] "eau".

- **L'état :** On distingue deux états pour les noms amazighes, l'état libre et l'état d'annexion.

L'état libre : dans ce cas, aucune modification n'est appliquée sur la voyelle initiale du nom. On dit que le nom est en état libre lorsqu'il s'agit : d'un mot isolé de tout contexte syntaxique (□□□□□ [atbir] "pigeon"), d'un complément d'objet direct (□□□□ □□□□ □ □□□□ [iTTfaslm g ufus] "Il tient un poisson à la main"), ou bien d'un complément de la particule prédictive □ [d] "c'est" (□ □□□□ [d aslm] "c'est un poisson").

L'état d'annexion : dans ce cas, une modification de l'initiale du nom dans des contextes syntaxiques déterminés est appliquée. Il prend l'une des formes suivantes : alternance vocalique □[a]/□[u] ou bien maintien de la voyelle initiale et ajout d'un □ [w] en cas des noms masculins à initiale □ [a] (□□□□□ [argaz] "homme" ->□□□□□ [urgaz]), addition d'un □ [w] pour ceux à initial □ [u] et d'un □ [y] aux noms à voyelle □ [i] (□□□ [ils] "langue" ->□□□□ [yils]). Pour les noms féminins, cet état est défini soit par la chute de la voyelle initiale (□□□□□□ [tamurt] "pays" → □□□□□ [tmurt]). L'état d'annexion se réalise dans les contextes syntaxiques suivants :

- Le sujet lexical suit le verbe =>ⵙⵎⵙ□ □□□□□□ → "Le professeur est venu"
- Après une préposition =>ⵙⵖⵍⵍⵉⵢⵍ□□□□□□ → "J'ai parlé au professeur"
- Après un coordonnant =>□□□□□ □ □□□□□□ → "la pluie et le froid"

3. 2. 3. 2 ***Verbe***

En amazighe, le verbe peut prendre deux formes : simple ou dérivée. Le verbe simple est composé d'une racine et d'un radical. Cependant le verbe dérivé est obtenu à partir des verbes simples par une préfixation de l'un des morphèmes suivants :□ [s]/□□ [ss], □□ [tt] et □ [m]/□□ [mm](Laabdelaoui, et al., 2012). La première forme correspond à la forme factitive, la deuxième marque la forme passive et la troisième désigne la forme réciproque. Le verbe, qu'il soit simple ou dérivé, se conjugue selon quatre thèmes :l'aoriste, l'inaccompli, l'accompli positif et l'accompli négatif, et possède deux modes : l'impérative et l'impérative intensive. Il est accompagné, en fonction des thèmes, de l'une des particules aspectuelles de l'amazighe.

*3. 2. 3. 3 **Particules***

En général, les particules sont les mots outils pour une langue donnée. Notons que les particules sont souvent des mots invariables, souvent assez courts, elles désignent un ensemble de mots amazighes qui ne sont ni des noms, ni des verbes, et jouent un rôle de désignateurs grammaticaux au sein d'une phrase. Cet ensemble comprend plusieurs formes, à savoir : les particules d'aspect, d'orientation et de négation; les pronoms indéfinis, démonstratifs, possessifs et interrogatifs; les pronoms personnels autonomes, affixes sujet, affixes d'objet direct et indirect, compléments du nom ordinaire et de parenté, compléments de prépositions; les adverbes de lieu, de temps, de quantité et de manière; les prépositions; les subordonnants et les conjonctions. Généralement, les particules sont invariables. Or, dans le cas de l'amazighe, similairement au cas du français, il existe des particules flexionnelles telles que les pronoms possessifs (□□□□□ "le sien" → □□□□□□ "le leur").

3. 3 Outils linguistiques pour la langue amazighe

Auparavant, la visibilité de la langue amazighe au Maroc était quasiment nulle. Récemment, et grâce aux revendications qui se sont faites à l'aide de l'IRCAM, elle a été soumise à un processus de codification et de standardisation, comme il a été déjà relevé. Une des préoccupations essentielles de l'IRCAM est de développer des outils linguistiques à l'égard de la promotion et la valorisation de la langue amazighe sur Internet et de la participation dans les processus multi-parties prenantes.

Il est à noter que la langue amazighe ne possède pas encore de ressources langagières et d'outils suffisants pour son traitement automatique. En plus, et face à l'augmentation vertigineuse des informations en langue amazighe, disponibles librement sur le Web, plusieurs recherches scientifiques ont été entamées dans ce sens, faute de les signaler tous, nous ne mentionnerons que les plus importants :

Ressources linguistiques :

- Les polices de caractères Tifinaghe[14,15](IRCAM, 2003; IRCAM, 2004)
- La création de corpus textuel amazighe (Boulaknadel, et al., 2011)et jeu d'étiquette (Ataa Allah, et al., 2014)

[14]Disponible: http://www.ircam.ma/fr/index.php?soc=telec&rd=1

[15]Disponible : http://www.ircam.ma/fr/index.php?soc=telec&rd=3

- Bases de données terminologiques (EL Azrak, et al., 2011; Frain, et al., 2014)
- L'élaboration de ressources d'apprentissage (Boulaknadel, 2016)
- Paradigmes flexionnelles (Raiss, et al., 2012; Nejme, et al., 2016)
- Corpus parallèle aligné par phrase (amazighe-anglais) (Miftah, et al., 2017)
- Dictionnaire multilingue amazighe-anglais-arabe-espagnol-français (Taghbalout, et al., 2017)

Outils pour le TAL :

- Correction orthographique (Es Saady, et al., 2009)
- Moteur de recherche (Ataa Allah, et al., 2010)
- Pseudo-racinisation (Ataa Allah, et al., 2010; Ataa Allah, et al., 2010)
- Concordancier(Boulaknadel, et al., 2010)
- Convertisseur (Ataa Allah, et al., 2011)
- Conjugueur (Ataa Allah, et al., 2014)
- L'étiquetage morphosyntaxique (Ataa Allah, et al., 2009; Outahajala, et al., 2014)
- Reconnaissance des entités nommées amazighes (Boulaknadel, et al., 2014; Talha, et al., 2018)
- Traduction automatique (Taghbalout, et al., 2016)
- La reconnaissance de la parole (Satori, et al., 2014; Elouahabi, et al., 2016)
- La reconnaissance optique des caractères Tifinaghes (Es Saady, et al., 2010; Fakir, et al., 2009; Amrouch, et al., 2010; Rabi, et al., 2017)

3. 4 Les défis de la langue amazighe en reconnaissance des entités nommées

La langue amazighe comme étant une langue peu dotée présente encore des défis qui sont bien évidement des éléments handicapants pour son développement en TAL, et notamment en REN amazighe qui est encore à l'état naissant et possède encore des difficultés non résolues et sa définition reste délicate en pratique. Certains défis principaux que peut affronter chaque concepteur de système de REN pour la langue amazighe sont les suivants :

- La langue amazighe est considérée comme une langue officielle du Maroc, paradoxalement, même avec l'existence du clavier et des supports amazighes et les

efforts de standardisation de l'orthographe, certains usagers de TIC continuent de faire des transcriptions en caractères latins ou arabes. Le phénomène de la latinisation et l'arabisation persiste.

- L'historicité de la littérature en langue amazighe est considérable, elle jouit d'une richesse lexicographique énorme. Cependant, étant une langue peu dotée, elle souffre d'un manque en termes de ressources linguistiques informatisées, de ressources dictionnairiques (de noms etc.), de répertoires toponymiques, de ressources langagières et outils du TAL.
- La REN amazighe est essentiellement confrontée aux problèmes de l'orthographe. En effet, la langue amazighe a une tradition littéraire très vigoureuse et diversifiée sauf que cette tradition n'a été que très rarement ancré par l'écrit. Les règles diffèrent selon le style d'écriture des mots, ceci peut se produire, à titre d'exemple, en utilisant ou en éliminant des espaces à l'intérieur ou entre les mots ([tadartino] [Tadartino] (ma maison)). Cette multiplicité de formes illustre la grande richesse orthographique de l'amazighe mais qui peut produire des problèmes de segmentations surtout en l'absence de standardisation.
- Bien que l'amazighe soit d'une tradition orale, mettre au point la grammaire de cette langue n'est pas chose aisée. Certes, toute langue, qu'elle soit écrite ou orale, dispose forcément d'une grammaire, la seule différence c'est que, dans le premier cas, l'établissement de cette grammaire est clair et, alors que dans le deuxième, elle est tout simplement compliqué.
- Similairement à d'autres langues naturelles, l'amazighe présente des incertitudes au niveau des classes grammaticales. En effet, la même forme convient à nombreuses catégories grammaticales, cela dépend du contexte dans la phrase. Par exemple, □□□□ [illi] peut être considéré comme verbe à l'accompli positif, il signifie «il existe », ou comme nom de parenté «ma fille».
- Les effets contextuels peuvent également donner à un mot une signification inattendue laissant cours à l'interprétation ou à la variation sémantique. L'amazighe, comme tout autres langues, souffre du problème des ambigüités qui se manifestent sous trois formes ambigüité lexicale, syntaxique et sémantique.
- L'absence de la distinction majuscule/minuscule : c'est un obstacle majeur pour la langue amazighe. En fait, la REN pour certaines langues comme les langues indo-

européennes se base principalement sur la présence des lettres majuscules qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin. Les lettres majuscules, néanmoins, ne se produisent pas, ni au début ni à l'initiale des noms propres amazighes.

- Il est un fait que la langue amazighe est non concaténative ayant une morphologie dérivationnelle et flexionnelle assez complexe et riche, les noms peuvent avoir plusieurs formes fléchies et dérivées, la simple suppression des suffixes ne peut suffire à regrouper des familles de mots. En effet, dans la pratique les affixes peuvent altérer le sens d'un mot.
- Les noms propres au niveau de la langue amazighe sont extrêmement nombreux, ont de nombreuses variantes ainsi qu'ils sont difficiles à détecter sans la présence d'un lexique, et c'est également le cas pour les noms d'organisations ou de produits, les noms de lieux même si ces derniers sont relativement stables par rapport aux précédents qui subissent fréquemment des changements.
- Le nombre des dialectes amazighes parlés au Maroc représente une source de richesse linguistique parce que cela fait apparaître une diversité culturelle et une dynamique assez spéciale. Cependant, cette multitude dialectale est aussi un problème du simple fait qu'elle est toujours présente dans un ensemble de ressources en dépit du processus de normalisation de l'orthographe.

3. 5 Conclusion

Dans ce chapitre, nous avons présenté un bref état des lieux de la langue amazighe ainsi que ses caractéristiques et son écriture Tifinaghe, nous avons ensuite donné une présentation succincte de sa morphologie. En outre, nous avons dressé un bilan de réalisation d'outils et de ressources linguistiques amazighes, permettant son traitement automatique et son intégration dans le domaine informatique. L'objectif était de faire une synthèse des travaux effectués pour la valorisation et la préservation de cette langue. Finalement, nous avons cité les difficultés entravant le traitement automatique de la langue amazighe et notamment la reconnaissance des entités nommées amazighes.

D'après notre analyse, nous remarquons que l'amazighe reste un champ d'investigation très large où beaucoup de recherches spécialisées restent à faire. Certes, les recherches menées ces dernières années ont fait prévaloir des attitudes positives envers la promotion de

l'amazighe. Néanmoins, les travaux réalisés jusqu'à présent, présentent un certain nombre de limites qui proviennent à la fois des caractéristiques de cette langue ainsi que la faible disponibilité des ressources de références (non standards) permettant son traitement. Dans le chapitre suivant, nous présentons nos contributions dans le domaine de la reconnaissance des entités nommées en langue amazighe en commençant par la création de notre propre corpus.

Chapitre 4 :
Implémentation des systèmes de reconnaissance des entités nommées amazighes

Le chapitre présent sera consacré à la description de la démarche suivie pour la construction de nos systèmes de REN amazighes, qui ont été implémentés en se basant sur différentes approches. La première méthode utilise une approche à base de règles basée sur un ensemble de règles et de lexiques créés manuellement. Dans la deuxième nous proposons une nouvelle version du système en utilisant une approche à base d'apprentissage qui repose sur l'utilisation d'un classifieur nommé « SVM » et d'un ensemble de caractéristiques. Nous terminons le chapitre par la présentation de notre troisième méthode basée sur la combinaison des deux approches précédentes et qui est une version améliorée des autres.

4. 1 Système à base de règles « RENAM » :

Le but principal de cette première contribution est de développer un système d'extraction et d'annotation d'entités nommées amazighes. Nous avons donc adopté, en premier temps, une approche qui fait tout d'abord intervenir un composant symbolique, c'est-à-dire un système d'analyse syntaxique à base de règles que nous avons créé spécifiquement pour la langue amazighe. Par ailleurs, cette approche nous a permis de valider concrètement la couverture et la pertinence de nos ressources créées. Cette méthode a déjà été étudiée et expérimentée dans nos différentes publications (Talha, et al., 2014; Talha, et al., 2014; Talha, et al., 2015; Boulaknadel, et al., 2014)et s'est révélée suffisamment efficace. Pour l'extraction des EN, nous avons fait appel à la plateforme GATE que nous allons présenter dans la section5. 1. 1 .

4. 1. 1 Architecture du Système :

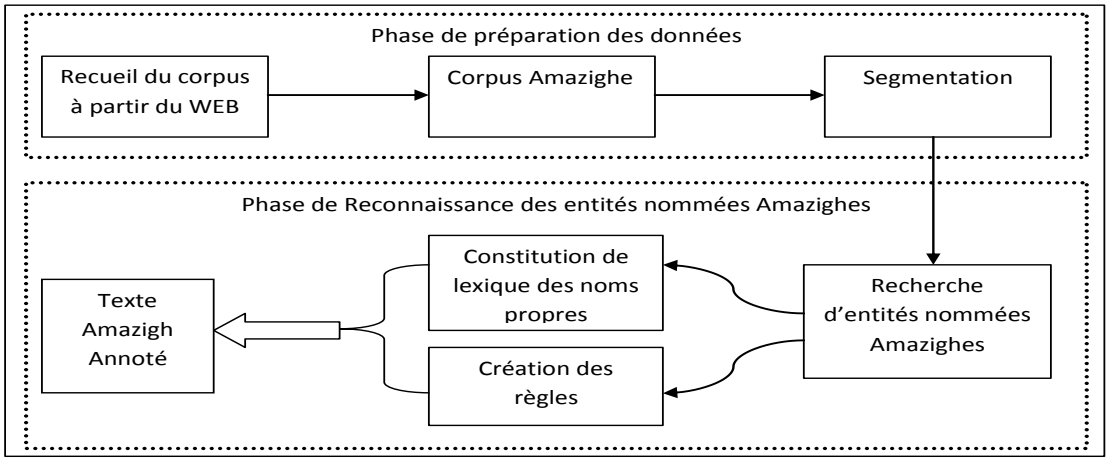

Figure 4-1 Architecture du système d'extraction des entités nommées amazighes (RENAM)

Le but de la chaîne de traitement que nous proposons dans laFigure 4-1 est, à partir d'un corpus donné en entrée, de générer en sortie un corpus annoté en termes d'entités nommées. Pour ce faire, une succession de traitements est nécessaire, nous les présentons succinctement ci-après, avant d'en fournir une description plus détaillée du module d'extraction d'entités nommées dans la suite de ce chapitre. Voici donc les cinq principales étapes de la chaîne de traitement en question (cf. Figure 4-2) :

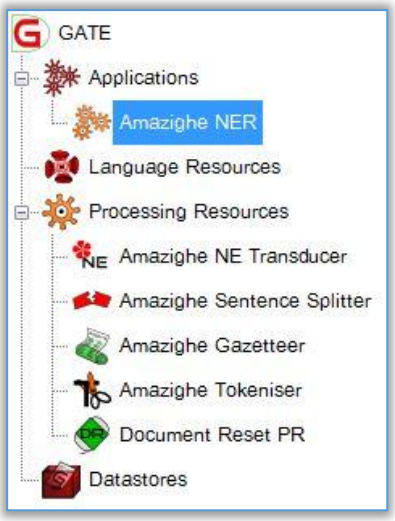

Figure 4-2 Amazighe NER - GATE

- Document Reset : supprime toutes les annotations précédentes mises sur le document.
- Amazighe Tokenizer(Séparateur de mots) : Dans l'analyse lexicale, la tokenisation (ou itémisation) est le processus de séparation d'un flux de texte en mots, symboles et d'autres éléments significatifs appelés items lexicaux ou tokens (ponctuations,

nombres, mots, etc.). Pour effectuer cette tâche, nous avons utilisé l'« Amazighe Tokenizer » qui procède à la segmentation d'un texte amazighe.

- Gazetteer : GATE utilise notamment un ensemble de listes ce qu'il appelle des gazetteers. Il cherche les éléments d'une liste prédéfinie annotée dans le texte et les annote en tant qu'entité « Lookup ». Nous allons, par exemple, avoir une liste contenant tous les noms de villes, ou de pays. Ces listes sont organisées en différentes catégories telles que « Location » qui se décompose en pays, ville, etc. Ces catégories vont nous servir lors de l'extraction des ENs.

- Sentence Splitter (Séparateur de phrase) : qui comme son nom l'indique, découpe le texte amazighe qui lui est fourni en entrée en un ensemble de phrases en fonction de la ponctuation. Nous présentons à titre d'exemple le texte suivant :

□□□□□□□ - □□□□ □□□ □ □□□□□□□ □□□□□□□ □□□□□□□ □□□□ □□□□, □□ □ □□□□ □□□□, □ □□□□□□□ □ □□□□□□□ □ □□□□□□□, □□□□□ □ □□□ □□□□□ □ □□□□□□□□□ □ □□□□, □□□□□ □□□□ □□□□□ □□□□□□□□□ □ □□□ □□□□□□□□ □□□□□□□□ □□□ □□□ □□□□ □□□ □ □□□□□□□ □□□□□□□ □ □□□□ □□□□□□□ □ □□□□□ □□□□□□□□□□□□ □□□□□□□.

Bamaku – ilkmbab n wadduragllidmuhmmdwisssdis, ad t insrrbbi, g tzdwit n usinas g bamaku, ammas n yanurzafitgdudant n mali, mayadigantiklitamzwarut n yattssutlttafrikitlli rad yawibab n wadduragllid s lkutdifwar d ghinyatakunakritdlgabun.

Bamako - Sa Majesté le Roi Mohammed VI, que Dieu L'assiste, est arrivé, mardi en fin d'après-midi à Bamako, pour une visite officielle en République du Mali, première étape d'une tournée africaine qui conduira le Souverain en Côte d'Ivoire, en Guinée Conakry et au Gabon.

- Named-Entitytransducer (NE Transducer) : C'est la partie de l'algorithme qui va utiliser toutes les informations précédentes pour localiser les entités nommées. En effet, GATE permet d'annoter un texte en se basant sur des critères et des règles que nous aurons établis manuellement. Ainsi, il est possible d'écrire des grammaires dont il se servira pour reconnaître toute entité présente dans un texte de référence qui correspondra aux règles fixées au préalable. Il convient de noter que nous avons créé nous-mêmes nos règles d'extraction qui permettent de repérer les entités nommées amazighes (Person, Organization, Location, Date, Money, Percent, Number.). Les règles utilisées sont écrites en JAPE (Java Annotation Patterns Engine).

Nous allons nous concentrer sur le module d'extraction, étant donné que la tâche d'extraction des EN amazighes se fait profondément lors de l'application de ce dernier.Il est divisé en deux phases nécessaires et complémentaires(Talha, et al., 2014; Talha, et al., 2014; Talha, et al., 2015; Boulaknadel, et al., 2014).

4. 1. 2 Constitution des listes des noms propres (gazetteers) :

Dans le but de pouvoir reconnaître automatiquement nos sept catégories des entités nommées, nous avons commencé, dans une première phase, par l'élaboration d'un lexique (gazetteers) à partir de différentes ressources librement disponibles sur internet. Nous avons aussi élaboré, des marqueurs lexicaux qui permettent de vérifier la présence des entités nommées en question. La constitution de ces lexiques électroniques est largement utilisée pour l'extraction des ENs. Enfin, nous avons enrichi ces lexiques, à partir du corpus et du Web, au fur et à mesure de nos tests. Lors de cette phase, notre système commence par la comparaison de chaque entrée dans le texte brut avec chacune des entrées des différents lexiques que nous avons construits.

4. 1. 2. 1 Lexique des « Noms de Personnes » :

Afin de construire les lexiques des entités nommées de type « Personne », nous avons compilé manuellement une liste qui comprend environ **3097**entrées de noms et prénoms amazighes ainsi que les noms et prénoms étrangers les plus courants transcrits en amazighe, provenant de plusieurs sites web. Nous avons pris en considération les prénoms simples (les prénoms ne contenant qu'un seul mot) et les prénoms composés qui peuvent comprendre jusqu'à quatre ou cinq mots (exp : Mohammed bin Zayed bin Sultan, □□□□□□□□ □□□ □□□□ □□□ □□□□□□). Cependant, notre lexique n'inclut qu'un nombre limité de noms composés qui sont majoritairement d'origine arabe.

4. 1. 2. 2 Lexique des « Noms de lieux / Localisation » :

Les noms de lieux en amazighe, comme dans d'autres langues, désignent tout ce qui représente une zone géographique, ce qui a été expliqué dans les travaux de (Piton, et al., 2004). Dans notre corpus, cette catégorie inclut tout ce qui représente les pays, les villes, les villages, les régions, les fleuves, les océans, les mers, les lieux sportifs (stades), les quartiers, les aéroports, etc. Ainsi, afin de créer le lexique des entités nommées de type « Localisation » ou encore appelé toponyme, nous avons collecté **3264** différentes entités à partir de

différentes sources principalement notre corpus ainsi que d'autres sites web (comme Wikipédia) couvrant diverses régions du monde entier.

4. 1. 2. 3 Lexique des « Noms des Organisations » :

A l'instar de la procédure d'élaboration de lexiques des deux entités nommées précédentes, l'identification des entités nommées de type « Organisation » a commencé par le développement d'un lexique qui contient **1022** noms d'organisation y compris les noms des Ecoles, les Instituts, les Partis Politiques, les Associations, les Ministères, les Banques, les Equipes sportives, les Hôpitaux, etc. Nous avons suivi la même démarche de collecte que celle employée avec les autres catégories, et qui est basée sur la collecte des données à partir des sites Web et de certaines ressources disponibles sur Internet.

4. 1. 2. 4 Lexique des « Expressions numériques » :

Pour les expressions numériques couvrant les différentes sous-catégories « Expressions Monétaires (exp : □□□□□, dolar, Dollar », « Pourcentages (exp : %) », « Les unités de mesure (poids, distance, volume, vitesse) (exp : □□□□□□ □□□□□□□,amitrumukââb, mètre cube) » et toutes autres expressions numériques, nous avons pu extraire et typer plus de **317** entités à partir des textes amazighes sur la base de leur sous-catégorie.

En observant les textes collectés, nous avons constaté que les chiffres amazighes employés sont systématiquement de cette manière : (1, 2, 3, 4, 5, 0, 7, 8, 9, 0).

4. 1. 2. 5 Lexique des « Expressions Temporelles » :

Concernant les entités nommées temporelles, elles incluent toutes les expressions employées pour fournir des indications temporelles sur la date (jour, mois, année/s, période), l'heure ou la saison de l'année et toute autre expression exprimant le temps. Dans notre lexique, nous avons pu dénombrer **221** entrées collectées, bien evidemment, à partir de notre corpus et aussi les manuels scolaires mis en ligne par l'IRCAM. Ces listes, transcrits en lettres amazighes, contiennent les noms des douze mois (exp : □□□□□, brayr, Février), les noms des jours de la semaine (exp : □□□□□, aynas, Lundi), les saisons (exp : □□□□□□, tagrst, Hiver) et aussi des expressions indiquant le temps (exp : □□□□, assa, aujourd'hui).

4. 1. 2. 6 Lexique de mots déclencheurs :

Ce dernier lexique contient les mots qui sont susceptibles d'apparaître dans le contexte et qui peuvent déterminer la présence des entités nommée en question (exemples : « □□□□□,

Lalla, Madame »,« ⵜⴰⵎⴷⵉⵏⵜ, tamdint, ville », « ⵜⴰⵎⴰⵡⴰⵙⵜ, tamawast, Ministère »), nous avons estimé un total de **522** entrées lexicales pour les différentes catégories.

Le Tableau 4-1recense les lexiques utilisés pour l'extraction des entités nommées amazighes, ainsi que les différents rôles que jouent chacune de ces listes.

Contenu du lexique	Rôles	Nombres d'entrées
Person_name.lst	Éléments de l'entité nommée « Personne»	3097
Title.lst	Marqueurs Lexicaux « Titres des civilité »	4
Titleh.lst	Marqueurs Lexicaux « Titre Honorifiques »	34
Titlem.lst	Marqueurs Lexicaux « Titres militaires »	8
Titlep.lst	Marqueurs Lexicaux « Titres politiques »	3
Titler.lst	Marqueurs Lexicaux « Titres religieux »	18
Jobtitles.lst	Marqueurs Lexicaux « Professions »	110
Organization.lst	Eléments de l'entité nommée « Organisation »	1022
Org_key.lst	Marqueurs Lexicaux de l'entité « Organisation »	98
Location.lst	Eléments de l'entité nommée « Localisation »	3264
Loc_prepo.lst	Marqueurs Lexicaux de l'entité « Localisation »	8
Loc_key.lst	Marqueurs Lexicaux de l'entité « Localisation »	144
Currency.lst	Éléments de l'entité nommée « Monnaie »	92
Percent.lst	Éléments de l'entité nommée « Pourcentages »	3
Numbers.lst	Éléments de l'entité nommée « Numéros »	222
Date.lst	Éléments de l'entité nommée « Date »	18
Date_key.lst	Marqueurs Lexicaux de l'entité « Date »	35
Day.lst	Éléments de l'entité nommée « Date »	26
Hour.lst	Éléments de l'entité nommée « Date »	39
Months.lst	Éléments de l'entité nommée « Date »	69
Time_key.lst	Marqueurs Lexicaux de l'entité « Date/Horaire »	60
Year.lst	Éléments de l'entité nommée « Date »	69

Tableau 4-1Lexiques développés pour la REN Amazighe

4. 1. 3 Développement des règles linguistiques

Une fois une EN reconnue grâce à un lexique, elle sera automatiquement retenue sans passer par la deuxième phase, qui est réservée exclusivement à la détection des EN ne figurant pas dans le lexique. Inévitablement, les lexiques qu'on a déjà construits ne sont pas aussi exhaustifs et certaines entités nommées peuvent ne pas y être présentes dans leur forme exacte. Il est difficile d'établir des listes finies et complètes de toutes les entités nommées car ces dernières se créent constamment, en outre tenir à jour ces listes est une tâche lourde et peu efficace. De ce fait, nous nous sommes tournés vers la prise en compte de la création des grammaires locales écrites à la main sous forme d'expressions régulières qui est amplement utilisée pour l'extraction des ENs. Cette méthode repose sur la présence des « marqueurs lexicaux » dans le contexte immédiat (droit ou gauche) d'une EN potentielle, comme a été

indiqué sur les travaux de (McDonald, 1996)dans le domaine de REN, qui consiste à déterminer les indices d'apparition des entités à extraire. La détection de ces indices s'est faite par observation des exemples mis à notre disposition dans le corpus AMCorp décrit dans la section 5.1et en repérant les points communs entourant une entité nommée donnée en tenant compte de la fréquence d'apparition de ces indices selon chaque type d'entité.

Ces indices peuvent être, d'une part, des indices internes, faisant partie intégrante de l'entité nommée, la grande majorité de ces indices sont localisé au début des entités. Comme dans les exemples suivants :

« □□□□□□□□ □□□□□□□□□□ □ □□□□□, [tarabbuttatunisit n bnzrt], Equipe tunisienne Bizertine »

« □□□□□□ □ □□□□□□□□ □ □□□□□□-□□□□□□□, [azaggaz n agraghlan n minara-Mrrakkch], L'aéroport international Marrakech-Menara »

D'une part, les indices externes permettent d'avoir une grande visibilité et plus d'informations sur les entités nommées en question, comme dans les exemples suivants :

« □□□□□□ □ □□□□□□, [Tmdint n Mrrakch], La ville de Marrakech »

« □□□□□□□□ □□□□□□□□□□ □□□□ □□□□□, [lkulunilmajurabd slam talal], Le colonel major Abd Slam Talal »

Ces indices sont des éléments textuels très importants à prendre en considération, surtout quand il s'agit de l'ambiguïté touchant l'assignation d'un type exacte aux ENs repérées. Par exemple, un nom de personne peut être confondu avec un nom de localisation comme dans le cas de la « □□□□□□□□□ □ □□□□□□ □□□□ □□□, [timzgida n hassanwiss sin], La mosquée Hassan II » où l'unité lexicale « □□□□□□ □□□□ □□□, [hassanwiss sin], Hassan II » n'est plus considéré comme un nom de personne, mais plutôt une partie intégrante d'entité de type localisation.

L'étude préalable sur le corpus AMCorp nous a permis dans un premier temps de recenser les différentes entités qu'il faut extraire et dans un deuxième temps d'étudier la structure syntaxique des différents types d'entités nommées ainsi que les marqueurs lexicaux. En effet, nous avons constaté que la plupart de ces entités nommées sont des termes composés de plusieurs mots surtout quand il s'agit des entités nommées de type « Organisation » et « Personne ».

Pour ce faire, nous avons établi manuellement un ensemble de règles pour extraire chacune des entités nommées en question. Le Tableau 4-2 suivant décrit le nombre de règles élaborées.

Entités Nommées	Nombre de Règles
Personne	28
Localisation	26
Organisation	18
Expressions Numériques	18
Expressions Temporelles	41
Total	131

Tableau 4-2 Tableau Récapitulatif des règles construites

Chaque règle d'extraction construite traduit une forme d'apparition d'une EN existante dans le corpus d'étude. Sachons que nous avons procédé par essais (et erreurs) successifs pour essayer de concevoir une règle adéquate. C'est-à-dire en rajoutant un terme à la règle lorsque la précision est faible, en enlevant un autre lorsque le rappel diminue.

Nous avons groupé les règles (écrites en JAPE) dans des fichiers séparés. Notons que pour savoir quelles règles faut-il appliquer lorsque plusieurs sont déclenchées parallèlement, nous utilisons un mécanisme de contrôle de priorités qui permettra la hiérarchisation de nos règles d'annotation et les appliquer selon un ordre prédéfini afin de produire des annotations valides. Nous avons priorisé les règles les plus longues et les plus complexes en premier, et nous avons laissé les règles les plus simples vers la fin, et ceci afin d'éviter de détecter partiellement des ENs.

Nous illustrons dans la suite quelques règles d'extraction des entités nommées en donnant leurs natures et des exemples appropriés.

*4. 1. 3. 1 **Règles de type « Personne » :***

Le fichier des règles pour l'extraction des entités de type personnes contient **28** règles. Nous illustrons dans ce qui suit un exemple d'une règle simple extraite du fichier des règles pour la catégorie personne.

```
Rule:   TitlePerson
Priority: 30
(TITLE)
(
 {Token.kind== word}
):person
-->
{
gate.AnnotationSet person = (gate.AnnotationSet)bindings.get("person");
gate.Annotation personAnn = (gate.Annotation)person.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
features.put("kind", "personName");
features.put("rule", "TitlePerson");
outputAS.add(person.firstNode(), person.lastNode(), "Person",
features);
}
```

Figure 4-3 Exemple 1 - "Règle d'extraction des entités nommées de type Personne "

La règle ci-dessus (cf.Figure 4-3) doit pouvoir extraire les entités nommées de type « Personne » comme celle en gras dans l'exemple qui suit :

□□□□□ □□□□□□□ / massa **tasaadit**/ « Madame**Tasaadit**»
<TitrePersonne><Nom>➔ Nom= « entité personne »

Dans ce cas de figure, la règle va réussir à reconnaitre les ENs qui sont précédés par les marqueurs lexicaux de type « Titre de civilités » décrits ici par l'expression « TITLE », et elle va assigner la catégorie « Personne » à ces entités reconnues. Notons que les titres précédant les noms de personnes sont exclus de cette catégorisation.

De façon similaire, nous avons construit les règles pour la détermination de l'entité personne fait usage à la présence :

- D'autres marqueurs de civilités à titre d'exemple (□□□□ (Mass, Monsieur), □□□□ (Mast, Mme), □□□□□ (Massa, Mme), □□□□□ (Lalla, Mme), …
- Les titres de toute sortes tels que :
  - Titres politiques : « □□□□□□ »[amawas], ministre
  - Titres honorifiques : « □□□□ⵒⵄⵄⵔ» [bab n waddur], Sa Majesté,
  - Titre religieux : « □□□ » [Chikh], Cheikh
  - Titres militaires : « □□□□□□□□ » [lkulunil], le colonnel

Prenons un autre exemple de règle permettant la reconnaissance d'une entité personne :

```
Rule:   RelationTwo
Priority: 50
(
  (
  {Token.string== "ⵄⴱⴷⵍ"}|
  {Token.string== "ⵄⴱⴷ"}|
  {Token.string== "ⵄⴱⴷⵍ"}|
  {Token.string== "ⵄⴱⴷⵍⵍ"}
  )
  {Token.kind== word}
):person
-->
{
gate.AnnotationSet person = (gate.AnnotationSet)bindings.get("person");
gate.Annotation personAnn = (gate.Annotation)person.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
features.put("kind", "personName");
features.put("rule", "RelationTwo");
outputAS.add(person.firstNode(), person.lastNode(), "Person",
features);
}
```

Figure 4-4 Exemple 2 - "Règle d'extraction des entités nommées de type Personne"

La règle ci-dessus (cf. Figure 4-4) permet l'extraction des noms de personnes comme celui de l'exemple suivant :

□□□ □□□□□□ / Abdllilah / AbdElilah

<□□□><Nom>➔ Nom Complet= « entité personne »

On constate ici que dans le cas où le contexte ne nous permet pas de reconnaitre une expression et que cette dernière n'est pas présente dans nos lexiques et est précédé par le mot clé « □□□, Abd » avec toutes ses différents formes d'écritures («ⵄⴱⴷⵍ, Abd », «ⵄⴱⴷⵍⵍ, Abdu », «□□□□, Abdu »), toute la séquence va être annoté comme étant une entité de type personne.

De la même manière, et pour donner suite à l'observation des autres prénoms composés présents dans notre corpus, nous avons créé une règle pour définir le lien de parenté, chaque culture a son propre système de nommage, prenons l'exemple de « □□□□□□ □□ □□□□ » (muhmmdbnzayd, Mohamed Ibn Zayd), le modèle suivant (cf.Figure 4-5) indique que quand

on rencontre l'une des combinaisons suivantes, toute la séquence peut être considéré comme un nom de personne :

<NomPropre><□□□(bnu, fils) ><NomPropre>➔Expression Complète= « entité personne ».

<NomPropre><□□□(ibn, fils) ><NomPropre>➔Expression Complète= « entité personne ».

<NomPropre><□□□(abu, père) ><NomPropre>➔Expression Complète= « entité personne ».

<NomPropre><□□(bn, fils) ><NomPropre>➔Expression Complète= « entité personne ».

<NomPropre><□□□□(ibnu, fils) ><NomPropre>➔Expression Complète= « entité personne ».

```
249 Rule:   RelationOne
250 Priority: 20
251 (
252   {Token.kind== word}
253   (
254   {Token.string== "ⵓⵏ"}|
255   {Token.string== "ⵜⵓⵏ"}|
256   {Token.string== "ⴰⵓⵏ"}|
257   {Token.string== "ⵓⵏ"}|
258   {Token.string== "ⵜⵓⵏ"}
259   )
260   {Token.kind== word}
261 ):person
262 -->
263 {
264 gate.AnnotationSet person = (gate.AnnotationSet)bindings.get("person");
265 gate.Annotation personAnn = (gate.Annotation)person.iterator().next();
266 gate.FeatureMap features = Factory.newFeatureMap();
267 features.put("kind", "personName");
268 features.put("rule", "RelationOne");
269 outputAS.add(person.firstNode(), person.lastNode(), "Person",
270 features);
271 }
```

Figure 4-5Exemple 3 - " Règle d'extraction des entités nommées de type Personne "

4. 1. 3. 2 Règles de type « Localisation » :

L'extraction des entités nommées de type « localisation » en amazighe est moins dure que l'extraction des entités de type « personne », car les noms des lieux qui représentent notamment les noms de pays et villes sont stables dans la mesure où les noms de lieux ne changent pas souvent. Quant aux autres noms de lieux, ils sont souvent associés aux mots déclencheurs qui aident à leur reconnaissance.

Par conséquent, les règles linguistiques pour la reconnaissance de ces entités sont écrites en se basant sur des préfixes de localisation et des prépositions qui accompagnent toujours les noms de lieux comme par exemple (‘□□□ (dar, en direction de)’, ‘□□ (ghr, vers)’, …).

Sauf que ces éléments ne s’avèrent pas toujours aussi pertinents, car sur la base de nos observations, nous avons constaté que ce type de règles peut générer des erreurs d’extraction. Prenons par exemple le cas des énoncés suivants :

« □□□□ □ □□□□□□□□□, [izdgh g lmansura], il habite à Al Mansoura »

« □ □□□□□□□□□ □ 08 □□□□□ 1945, aux manifestations du 08 Mai 1945 »

On remarque que la préposition « □, g » peut précéder des noms de lieux comme elle peut précéder d'autres catégories. La présence de cette préposition n’implique pas forcément que le nom qui vient ensuite est un nom de lieu, quand on a appliqué la règle ça nous a généré beaucoup d’erreurs d’extraction, nous étions obligés d’ajouter d’autre critères d’extraction précédant ces prépositions afin d’augmenter la précision de notre système. Nous nous sommes servis d’une liste de verbes exprimant une localisation afin de renforcer l’extraction des ENs en question.

En outre, nous avons construit une liste des mots déclencheurs comme par exemple, « □□□□□□ » [tmdint], ville. Ces marqueurs lexicaux sont exploités pour la description des règles de reconnaissance des entités de type « localisation ».

Voici un exemple présenté dans la Figure 4-6 décrivant une règle pour l’entité Localisation. Prenons le cas de « ⴰⵙⵉⴼ ⵏⵙⴱⵓ,[Asif n sbu], rivière de Sebou », ce qui signifie que tout mot qui vient après un marqueur lexical de type localisation représenté ici par l’expression (LocationKey) et qui est dans ce cas « □ □ □ □,asif, rivière » sera annoté comme étant un nom de lieu.

```
Rule:   LocationkeyBisBiiis
Priority: 30
(
  (LocationKey)
  {Token.string== "ⵏ"}
  (LocationName)
):location
-->
{
gate.AnnotationSet location = (gate.AnnotationSet)bindings.get("location");
gate.Annotation locationAnn = (gate.Annotation)location.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
features.put("kind", "locationName");
features.put("rule", "LocationkeyBisBiiis");
outputAS.add(location.firstNode(), location.lastNode(), "Location",
features);
}
```

Figure 4-6Exemple - " Règle d'extraction des entités nommées de type Lieux "

4. 1. 3. 3 Règles de type « Organisation » :

Contrairement aux noms des lieux, les noms des organisations sont assez nombreux et ne sont pas aussi facilement quantifiables, étant donné que leur apparition ou leur disparition dépend de la situation dans le monde. Par ailleurs, il arrive aussi qu'une organisation puisse s'écrire de différentes manières, soit sous une forme longue ou une forme courte. Par exemple : « □□□□□□□□ □□□□□□□□ □ □□□□□□ □□□□,[tamsmunt n tamatayt n timttaimun],l'Assemblée générale des Nations Unies » qui est une forme longue, peut apparaître dans un autre texte avec une forme plus réduite comme « □□□□□□ □□□□,[timttaimun], les Nations Unies », et d'après nos observations ces mots d'organisations n'obéissent pas à des règles strictes, ainsi la structure des mots d'organisations n'est pas standard, n'importe quel mot peut faire partie de cette catégorie. Néanmoins, dans la majorité des cas qu'on a relevé, les noms d'organisations sont formés de noms communs, des noms propres, d'adjectifs et des nationalités. En revanche, nous avons constaté l'insertion des mots déclencheurs en amazighe qui font partie de la composition de l'entité nommée « organisation » ce qui nous amène à créer une liste de règles de reconnaissance en fonction de ces déclencheurs.

On peut donc reconnaître l'EN « □□□□□□□□ □ □□□□□, [tamawast n tdusi], Ministère de la Santé » grâce à la présence du déclencheur « □□□□□□□□, tamawast »,qui constitue un indice interne car il fait partie de l'EN « organisation » elle-même. De manière similaire, on reconnait que « □□□□ □ □□□□□□, [bank n lmghrib], Bank Al-Maghrib » est une EN de type « Organisation », dans ce cas on ne va pas annoter □□□□□□ comme étant nom de lieu.

A partir de ces deux petits exemples, on peut déduire qu'un mot déclencheur d'organisation suivi d'un nom de lieux ou un mot inconnu, implique la catégorisation de toute l'entité comme étant EN de type organisation. Ceci se traduit par la règle suivante (cf. Figure 4-7):

```
Rule:   OrganizationkeyCountry
Priority: 30
(
  (Organizationkey)
  {Token.string== "|"}
  (CountryName)
):organization
-->
{
gate.AnnotationSet organization = (gate.AnnotationSet)bindings.get("organization");
gate.Annotation organizationAnn = (gate.Annotation)organization.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
features.put("kind", "organizationName");
features.put("rule", "OrganizationkeyCountry");
outputAS.add(organization.firstNode(), organization.lastNode(), "Organization",
features);
}
```

Figure 4-7Exemple - " Règle d'extraction des entités nommées de type Organisations "

L'expression {OrganizationKey} fait appel à la liste des marqueurs lexicaux indiquant la présence d'un nom d'organisation tandis que, l'expression {CountryName} fait appel à la liste des noms de lieux présents dans notre lexique. Dans ce cas, nous serions amenés à considérer l'association suivante :

<OrganizationKey><□><CountryName>➔ Expression Complète= « entité organisation ».

4. 1. 3. 4 Règles de type « Expressions numériques » :

L'utilisation dans la langue amazighe des expressions de mesures, de monnaies ou des pourcentages suit une structure et des règles d'écriture bien établies. Cette approche facilite la formulation des règles d'extraction pour cette catégorie d'entités nommées.

Nous nous sommes servis des listes qu'on a créées précédemment, contenant un ensemble de déclencheurs de cette catégorie pour la formulation de nos règles d'extraction de ces entités. Dans ce qui suit, nous illustrerons un exemple de règle qui permet de localiser une entité numérique (cf. Figure 4-8).

```
371  Rule:   MoneyCurrencyUnitAmzone
372
373    (
374        (AMOUNT_NUMBER)
375        ({Lookup.majorType == number})
376        ({Token.string == "|"})
377        ({Lookup.majorType == currency_unit})
378    )
379  :number -->
380    :number.Money = {kind = "number", rule = "MoneyCurrencyUnitAmzone"}
```

Figure 4-8 Exemple - " Règle d'extraction des entités nommées de type Numérique"

Dans la règle ci-dessus, l'expression Macro (AMOUNT_NUMBER) représente tous les nombres décimaux séparés par une virgule ou un point, l'expression ({Lookup.majorType == number}) permet de repérer un nombre (en chiffre ou en lettre) précédant l'unité monétaire et qui sont présents dans nos lexiques, tandis que l'expression ({Lookup.majorType == currency_unit}) renvoi à la liste d'expressions de monnaie que nous avons préalablement compilée.

Cette règle simple permet de localiser systématiquement des expressions comme : « 2.5 □□□□□□ □ □□□□□,[2.5 umlyar n ddrhm], 2.5 milliards de dirhams ».

*4. 1. 3. 5 **Règles de type « Expressions Temporelles » :***

Nous nous sommes appuyés sur notre corpus de travail pour saisir que la majorité des dates sont exprimées sous forme de chiffres et parfois en toutes lettres (p. ex :□□□ □ □□□□□ 10 □□□□□□ 2015, [ass n ukras 10 yulyuz 2016], « Jour du » mercredi 10 juillet 2015).

A noter qu'il y a plusieurs formats variables pour les dates, à titre d'exemple, la date « 16-05-2016 » est différente de « 16 □□□□□ 2016 » au niveau orthographique, même si elles ont la même valeur sémantique. Les deux formats d'écriture ont le même sens. C'est pour cela nous

avons créé une liste de règles permettant de couvrir le maximum possibledes différentes formats présentes dans notre corpus.

La règle suivante (cf. Figure 4-9) permet de détecter les entités temporelles qui s'écrivent sous cette forme : « 29 □□□□□□□□ □□7 □□□□□□□□, [29 nuwanbirar 7 dujanbir], 29 Novembre au 7 Décembre » ou encore « □□□□□ □ □□□ □□□□□□□□ □□ □□ □□□□□□□□, [agnar d tzanuwanbirar sa dujanbir], vingt-neuf Novembre au sept Décembre ».

```
Rule:   DateNumaar

(
((NUM) | ({Lookup.majorType == number}))
(MONTH_NAME)
({Token.string== "ⴰO"})
(NUM | ({Lookup.majorType == number}))
(MONTH_NAME)
)
:date
-->
 :date.Date = {kind = "date", rule = "DateNumaar"}
```

Figure 4-9Exemple - " Règle d'extraction des entités nommées de type Temporelle"

L'expression ((NUM) | ({Lookup.majorType == number})) permet de détecter les nombres « (29 ou □□□□□ □ □□□, [agnar d tza], vingt-neuf), et (7 ou □□, [sa],Sept) », alors que (MONTH_NAME) va parcourir le lexique des mois pour détecter « Novembre □□□□□□□□, [Nuwanbir]et Décembre □□□□□□□□, [Dujanbir] ». On peut donc déduire que toute forme écrite comme suit, doit être considéré comme entité temporelle :

< Numéro (chiffre ou lettre) ><Mois ><□□>< Numéro (chiffre ou lettre) ><Mois >➔ Expression Complète = « entité temporelle ».

4. 2 Système à base d'apprentissage :

Le but principal de cette deuxième contribution (Talha, et al., 2018) est de réaliser un système de reconnaissances des entités nommées amazighes en intégrant une approche par apprentissage supervisée.

4. 2. 1 Architecture du système

Son architecture présentée à la Figure 4-10, s'articule autour de quatre étapes essentielles, brièvement décrites dans ce qui suit, qui effectuent un traitement séquentiel immédiat des données : prétraitement lexical, phase d'apprentissage (y compris l'extraction des descripteurs (features)) et ensuite la reconnaissance des entités nommées. L'entrée de ce processus est un ensemble de textes bruts amazighes, la sortie est une représentation sémantique de ce dernier en répondant à la question d'extraction des entités nommées.

Prétraitement lexical : Il s'agit d'un processus qui comprend deux opérations suivantes : segmentation du texte en phrases et en unités lexicales.

Étape d'apprentissage : Dans cette étape, un ensemble de caractéristiques est utilisé afin de représenter d'une manière générale les tokens. Elle Consiste à prendre en entrée les vecteurs de caractéristiques, le système apprend sur ses vecteurs puis prédit les classes d'une nouvelle entité selon le modèle de classification.

Phase de test : Cette étape exploite le fruit de la précédente pour annoter et classer définitivement les entités nommées amazighes présentes dans notre corpus de test.

Il faut noter que nous avons utilisé le classifieur SVMlightquia la particularité d'être directement utilisable depuis la plateforme GATE.

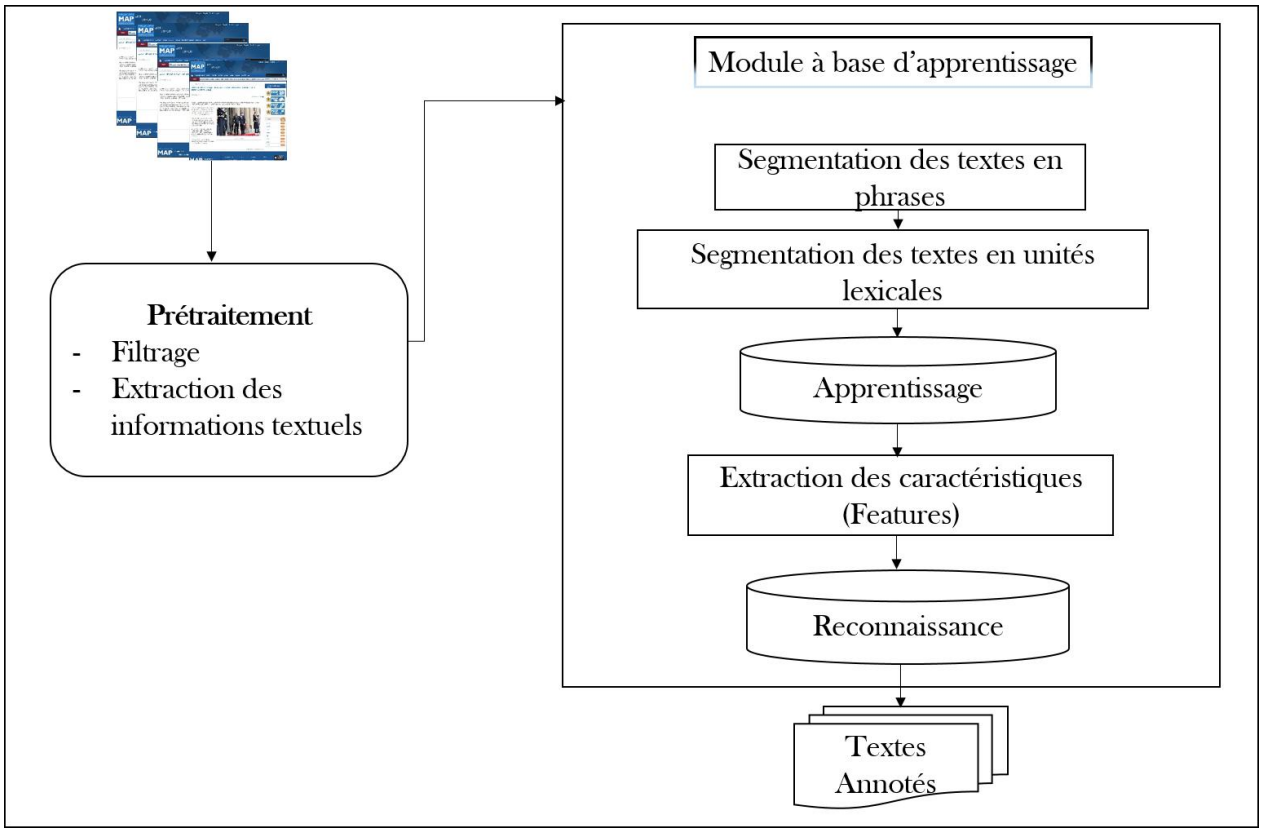

Figure 4-10 Architecture de notre système d'extraction des entités nommées amazighes en utilisant la méthode SVM

4. 2. 2 Machine à Vecteur Support:

Les « Supports Vectors Machines » (SVM) appelés aussi « Maximum Margin Classifier », « machines à vecteurs supports » ou « séparateurs à vastes marges », sont des techniques d'apprentissage supervisé basées sur la théorie de l'apprentissage statistique ou automatique.Ces techniques sont largement exploitées et elles ont fait preuve à résoudre des problèmes de discrimination (prédiction d'appartenance à des groupes prédéfinis) et de régression (analyse de la relation d'une variable par rapport à d'autres) (Silva, et al., 2009; Joachims, 1998). Les SVM sont apparus en 1995 suite aux travaux de Vapnik(Vapnik, 1995; Vapnik, et al., 1995). Ils sont à la base des classifieurs discriminants à deux classes, autrement dit, ils visent à séparer les données étiquetées en deux sous-ensembles disjoints, le principe de base (comme illustré dans la Figure 4-11) repose donc sur une analyse mathématique précise et avancée du problème de l'apprentissage qui cherche à calculer l'hyperplan séparateur optimal qui sépare le mieux un espace vectoriel en classes (si le problème est linéairement séparable). Le séparateur est disposé de manière à maximiser la

marge qui est la distance entre la frontière de séparation et les échantillons les plus proches de chaque classe qui sont appelés vecteurs supports.

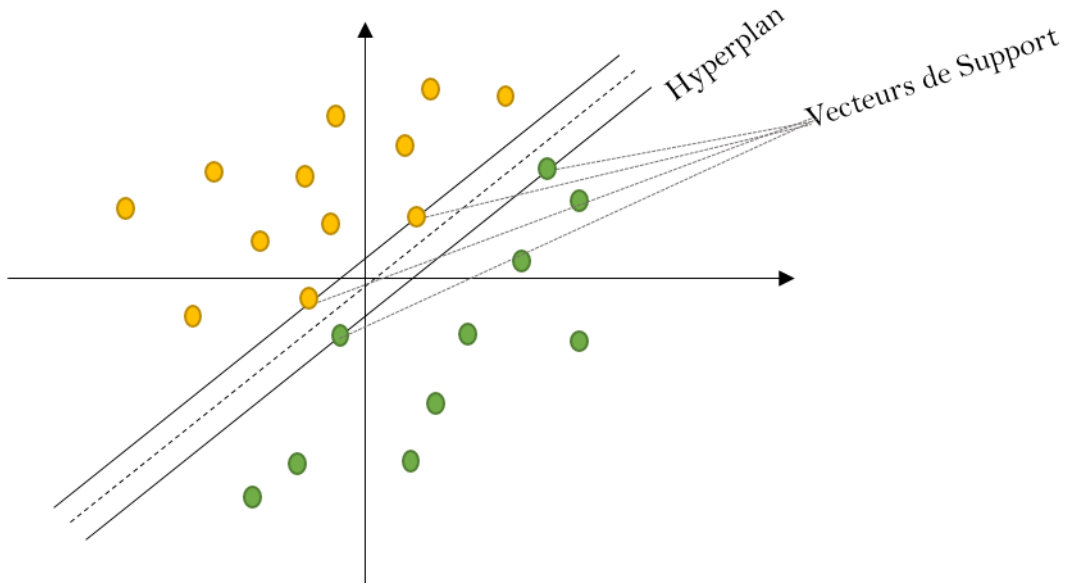

Figure 4-11 Exemple d'un hyperplan optimal et les vecteurs supports

En revanche, lorsque le problème est non linéairement séparable ou bien lorsqu'il contient des données bruitées, des fonctions noyaux (kernel en anglais) sont ajoutés afin de remédier à ce problème et de le rendre linéairement séparable. Les SVM devient capable de considérer le problème dans un espace de dimension plus élevé dans lequel il existe probablement un séparateur linéaire. Ce mécanisme de projection permet de changer l'espace pour réaliser l'apprentissage, c'est ce qui fait véritablement la force des SVM. Ce phénomène est illustré dans la Figure 4-12:

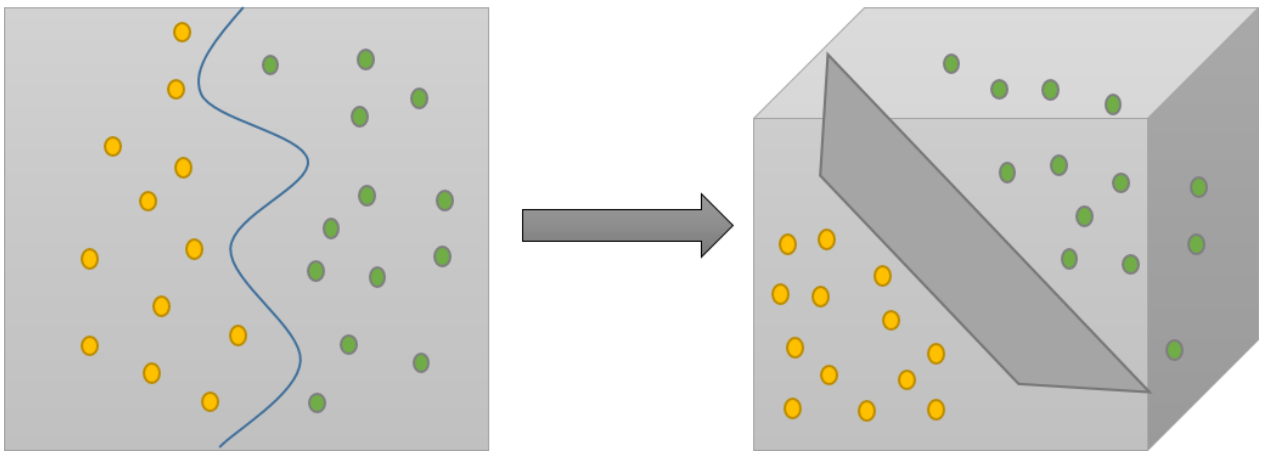

Figure 4-12 Transformation des données dans un espace de grande dimension

Il faut noter qu'il existe plusieurs types de noyaux utiles mais seulement les grands classiques noyaux linéaire et polynomial sont supportés par la plateforme utilisée (GATE).

Comme indiqué ci-dessus les SVM présentent une solution au problème de classification binaire. Or, dans notre cas, pour réaliser la reconnaissance des entités nommées amazighes, on a plusieurs classes. Nous faisons appel donc à SVM multi-classes qui est une extension de SVM. Deux méthodes principales sont proposées (Duan, et al., 2005):

Un-contre-tous (en anglais One-vs-All) : Cette méthode peut être considérée comme une généralisation du cas binaire pour traiter le cas de la classification multi-classe, elle est probablement la plus ancienne et qui consiste tout simplement à transformer le problème de T classes en T classifieurs binaires. Elle détermine pour chaque classe T un hyperplan qui la sépare de toutes les autres classes. L'apprentissage d'un classifieur T_N à vecteurs supports s'effectue en considérant tous les exemples de la T_N classe dans la région positive en leurs attribuant le label (+1) et tous les autres exemples dans la région négative en leurs attribuant le label (-1). En fait, cela veut dire que pour 5 classes nous aurons 5 classifieurs, et que chaque classifieur est un classifieur binaire entrainé sur tout le corpus étiqueté. En phase de test, le classifieur qui répond le mieux, donnant la valeur de confiance (e.g. la marge) la plus élevée, c'est celui qui remporte le vote. Le résultat est constitué par une liste ordonnée des classes, basée sur les scores obtenus. Cette méthode repose sur le principe de "a winner-takes-all, le gagnant emporte le tout".

One-vs-One (un contre un) : dans ce cas le problème est transformé en T (T-1) / 2 classifieurs binaires, T représente le nombre de catégories ou classes et i = 1, . . ., T représente les classes, et que chaque classe i est comparé à chaque classe j, i étant une classe positive et j une classe négative (i #j). Chaque classifieur est alors entraîné à différencier deux classes spécifiques. Concrètement, cela signifie que pour 5 classes nous aurons 10 = (5*4/2) paires de classes et donc 10 classifieurs. Cette méthode discrimine chaque classe i de chaque autre classe j. En phase de test, la classification d'un nouvel exemple se fait par un vote majoritaire qui permet de déterminer sa classe. Pratiquement, quand un nouvel exemple doit être classifié, celui-ci est soumis à l'ensemble des 10 classifieurs.

4. 2. 3 Configurations : Sélection des descripteurs

La qualité d'un système de reconnaissance des entités nommées repose fortement sur les données qui lui sont fournies en apprentissage, et notamment sur les descripteurs linguistiques (encore nommés traits ou attributs) décrivant les unités à classe. Ils consistent en

des éléments d'information qui seront associés à chaque token du texte. Le choix de ces descripteurs peut se révéler décisif. Nous décrivons ci-dessous les principaux que nous avons sélectionnés afin de réussir notre système :

- **Descripteur issu de Gazetteers :** ce descripteur est le plus privilégié vu sa forte pertinence dans notre système de reconnaissance des entités nommées amazighes. Pour chaque token w, il cherche s'il figure dans l'une des gazetteers que nous avons déjà créées (personne, localisation, organisation, etc.) (cf.section 4. 1. 2) ainsi il vérifie si le token (2 voisins gauches / 2 voisins de droites du token courant) appartient aux mêmes Gazetteers. Ce descripteur permet de révéler le contexte des entités nommées.

- **Informations contextuelles :** ce descripteur a une très grande importance, dans notre démarche. Nous avons opté pour l'utilisation d'une fenêtre contextuelle de taille 5 centrée sur le mot courant, c'est-à-dire que l'étiquette courante est estimée à partir des tokens et de leurs traits situés dans la fenêtre locale [−2, +2] entourant la position courante 0 de décision.

- **Ponctuation et nombres :** sur le plan lexical, des descripteurs de base ont été utilisé, ces traits testent respectivement la présence de caractères de ponctuations et de chiffres dans le mot courant ainsi que dans les deux mots gauches voisins et les deux mots droits voisins. Ces deux descripteurs sont notamment utiles pour mettre en évidence certaines entités numériques (exp. dates avec un séparateur généralement constitué d'une ponctuation).

- **La longueur du mot :** elle fait partie des descripteurs orthographiques, elle désigne le nombre de caractères dans le mot. Ce descripteur est envisagé pour distinguer les mots outils (composés d'un nombre réduit de caractères) des autres mots (potentiellement des entités à extraire), ceci se fait en partant du constat que les mots très courts représentent rarement des entités nommées.

- **Type de Token:** le modèle est construit uniquement sur la base des tokens tels qu'ils se présentent dans les textes sans recourir à d'autres types d'information. Le but principal est de définir si le token est un mot, un nombre, une ponctuation ou une autre unité de texte.

4. 3 Système Hybride :

Dans la lignée de nos travaux menés précédemment, nous avons conçu une extension des systèmes de reconnaissance des entités nommées amazighes(Talha, et al., 2015; Talha, et al., 2018)consistant en l'addition au système à base de règle initial d'un mécanisme d'apprentissage automatique afin de maximiser les performances de notre système hybride. Nous allons vérifier que cette technique d'hybridation est satisfaisante face aux autres possibilités.

Grâce à cette approche hybride, nous pouvons exploiter les forces de chacune des autres approches afin de couvrir un grand nombre de cas de figure problématiques liés à la reconnaissance des entités nommées amazighes. Le système hybride que nous proposons repose sur une mise en séquence (détaillée dans la Figure 4-13) de trois principaux modules à savoir : un module permettant le prétraitement lexical, première reconnaissance des EN grâce au module à base de règles, apprentissage et seconde reconnaissance des EN grâce au module à base d'apprentissage : chaque méthode utilise en entrée les informations données en sortie de l'autre méthode.

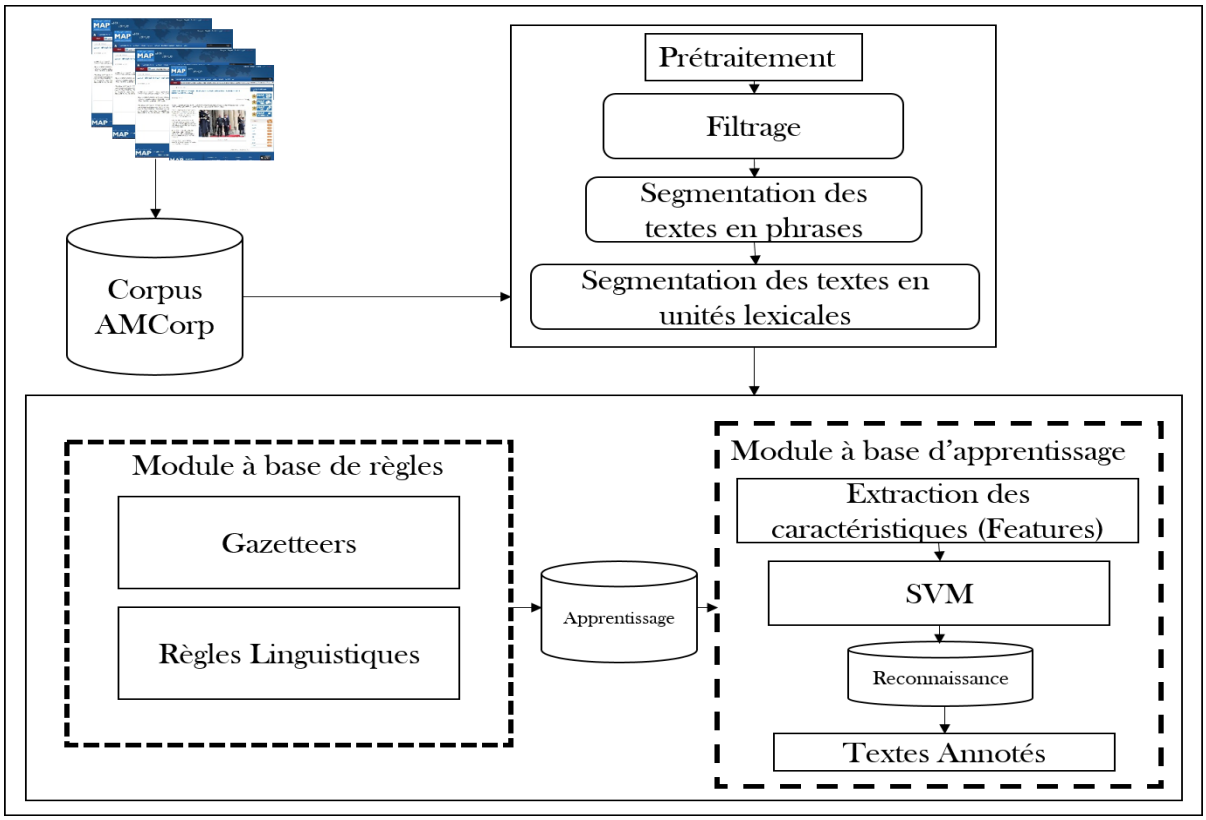

Figure 4-13 Architecture de notre système d'extraction des entités nommées amazighes en utilisant une approche hybride

La première reconnaissance se fait par le module à base de règles qui est une reproduction du système RENAM (cf. section4. 1). Il est construit à partir d'un ensemble de règles linguistiques et un ensemble de gazetteers.

Le module à base d'apprentissage exploite le fruit de cette acquisition pour apprendre et annoter cette fois-ci définitivement les entités nommées amazighes dans un texte, en considérant que les ENs annotées par notre module à base de règles sont correctes et n'utilisant le classifieur SVM que pour prédire les autres qui n'ont pas pu être annotées.Le SVM dans ce cas apprend la structure et le contexte dans lequel apparaissent les EN.

Il est à noter que ce module repose sur la sortie du système à base de règles, les descripteurs et l'algorithme de classification (cf. section 4. 2. 2). Les descripteurs étudiés dans notre système hybride sont divisés en : descripteurs dérivés du module à base de règles, les descripteurs des gazetteers, descripteurs contextuels, longueur des mots, type de tokens et les ponctuations.

La sortie du système hybride de REN, à base de SVM, est analysée et exploitée afin d'améliorer le module.

Ce système peut être qualifié d'hybride dans la mesure où les entités ne sont pas directement acquises par apprentissage à partir d'un corpus, mais le système possède généralement un ensemble de gazetteers et un ensemble de règles initiales.

4. 4 Conclusion

Dans ce chapitre, nous avons proposé trois systèmes de reconnaissance permettant d'identifier dans des textes amazighes de nature journalistique des entités nommées et plus particulièrement les noms de personnes, les noms de lieux, les noms d'organisation, les expressions temporelles et les expressions numériques, fondés sur trois approches à savoir celle à base de règles, statistique supervisée et hybride.

Nous avons commencé par la présentation du système RENAM qui utilise une approche **à base de règle**, la construction de ce dernier est composée de deux étapes primordiales : Constitution des lexiques des entités nommées et la création des règles d'extraction. Le processus de reconnaissance à travers ces règles fait aussi appel aux entrées disponibles dans les lexiques. Nous avons élaboré au total **8443**entrées de gazetteers et **131** règles d'extraction des entités nommées.

Nous avons ensuite présenté l'architecture de notre système de reconnaissance des entités nommées pour la langue amazighe reposant sur une approche **à base d'apprentissage**. La particularité de ce système réside dans l'exploitation d'un corpus d'apprentissage à partir duquel les entités nommées visées ont été annotées, l'objectif est d'« apprendre » des traits (ou caractéristiques) de nature linguistique et statistique caractérisant ces entités afin d'en déduire des modèles statistiques de reconnaissance et de classification. Nous avons ensuite introduit d'une manière simple et complète donnant une vision générale du concept des « Support Vecteur Machine » qui sont grandement utilisés en la reconnaissance des entités nommées ainsi que les vecteurs de caractéristiques que nous avons utilisés.

Partant de ces systèmes, nous avons étudié les possibilités de combiner les deux approches citées auparavant, nous avons pu conçu un système reposant sur une approche **hybride**. Un point important à noter est que ce système se base sur un module à base de règles renforcé par un lexique des entités nommées amazighes et un ensemble de règles linguistiques, alors que le module par apprentissage est utilisé pour supporter le premier.

Chapitre 5 :
Présentation et Analyse des résultats

Après avoir détaillé les systèmes proposés pour la reconnaissance des entités nommées amazighes, nous présentons dans cechapitre la plateforme utilisée afin d'expérimenter et valider nos systèmes. En effet nous avons choisi d'utiliser GATE, un environnement de développement conçu spécialement pour l'ingénierie des langues, pour mettre en œuvre les différents systèmes que nous avons proposé afin de reconnaitre les entités nommées amazighes.

Ensuite, nous présentons en particulier les propriétés fondamentales et la méthode principale qui a conduit à la constitution de notre propre corpus amazighe « AMCorp » manuellement annoté en entités nommées en se basant approximativement sur le schéma d'annotation de MUC-7.

L'évaluation des systèmes conçus est effectuée par l'utilisation des métriques d'évaluation : Rappel, Précision et F-mesure. Ensuite, une étude comparative des trois systèmes est effectuée. Enfin, en guise d'une discussion, nous présentons et nous interprétons les résultats obtenus en appliquant successivement les systèmes proposés précédemment lors de l'étape d'évaluation, à savoir : système à base de règles, celui à base d'apprentissage et système hybride.

5. 1 Protocol Expérimental :

Les plateformes linguistiques sont des environnements permettant d'implémenter et de concrétiser des travaux modélisés manuellement. Elles fournissent un ensemble de traitements sous format graphique. Elles permettent également de construire de nouvelles ressources où d'exploiter les anciennes afin de réaliser une tâche précise. Dans ce qui suit, nous présentons la plateforme linguistique utilisée « GATE » et les ressources implémentées sur cette plateforme ainsi que la typologie des entités nommées adoptée.

5. 1. 1 Aperçu général sur GATE

GATE[16] (General Architecture for Text Engineering) est une plateforme open source, conçue spécialement pour l'ingénierie textuelle(Cunningham, 2000; Gaizauskas, et al., 1996; Cunningham, 2002). C'est une infrastructure de développement de traitement du langage humain développée par les chercheurs de l'Université de Sheffield (Royaume-Uni) et est exploitée dans une vaste variété de recherche et de projets de développement incluant l'extraction de connaissances pour l'anglais, l'espagnol, le chinois, l'arabe, le français, l'allemand, l'hindi, le cebuano, le roumain et le russe.

GATE nous est apparu bien adapté au développement de notre système car il permet de fournir un Framework qui puisse implémenter l'architecture et être utilisé pour l'exploitation des traitements linguistiques dans diverses applications. En outre, il dispose d'une grande communauté d'utilisateurs qui permettra de disposer d'un ensemble de solutions d'aide et de support (forum, liste de diffusion, tutoriels, etc.), point indispensable lorsque l'on débute avec un tel outil et quand on l'applique pour une nouvelle langue.

Par ailleurs, la plateforme est basée principalement sur le principe d'une chaîne de traitements(« pipeline ») composée de plusieurs modules indépendants (nommés « ProcessingResources » PR) dédiés à l'analyse textuelle appliqués successivement sur un ou plusieurs textes (nommés « LanguageResources » LR). Ces modules servent aux tâches de Tokenisation de textes, Segmentation en phrases, Étiquetage morphosyntaxique (Part Of Speech Taggers), Reconnaissance des Entités Nommées, etc. Ces différents modules annotent chacun à leur tour le texte puis le document annoté est retourné à l'utilisateur au format XML. Il faut aussi noter que GATE dispose d'une interface graphique qui permet de charger de nouveaux plugins et ressources, de les paramétrer et de les combiner au sein d'une même chaîne de traitement.

GATE dispose d'une interface graphique qui permet d'annoter manuellement un ensemble de textes afin de créer un corpus de référence. Elle permet aussi de visualiser les résultats d'extraction. Le module d'extraction des entités nommées proposé dans GATE nommé ANNIE (A Nearly-New Information Extraction System) (Cunningham, et al., 2002) est réalisé selon une approche symbolique basé sur le formalisme **JAPE** (Java Annotation Patterns Engine) (Cunningham, et al., 1999)qui est un transducteur à états finis permettant de

[16]http://gate.ac.uk

définir les contextes d'apparition des unités à extraire dans le but de les repérer et les annoter. Cependant, ANNIE est conçue pour analyser les textes journalistiques en langue anglaise et dédiée à l'extraction des entités nommées pour l'anglais. Pour cela, nous avons dû réécrire entièrement toutes les règles de reconnaissance et nous avons créé notre propre lexique. Grosso modo nous avons conçu notre propre plugin nommé « NER Amazighe » qui nous a permis de procéder à la reconnaissance des entités en langue amazighe.

5. 1. 2 Préparation du corpus

Il est intéressant de collecter un nombre suffisant de textes qui vont servir de base à la mise en place de notre système de reconnaissance des entités nommées amazighes. La collection de ces textes va servir de corpus d'observation et d'analyse (afin de construire les règles) et de corpus de test (afin de procéder à la reconnaissance de ces entités).

Pour mener à bien ce travail et couvrir les différents thèmes, il nous a fallu recueillir divers articles et constituer manuellement un nouveau corpus de référence hétérogène et exploitable pour la création de notre système. En outre, notre second objectif est d'offrir à la communauté des chercheurs en TAL amazighe, ainsi qu'à un public averti, une ressource linguistique à partir de laquelle il sera possible d'envisager toutes autres sortes d'exploitation, et qui sera le fondement à partir duquel des traitements plus élaborés dans divers domaines peuvent être construits, comme dans : La traduction automatique, les systèmes Questions-Réponses, etc. Enfin, de nombreuses conférences dont l'objet est celui de l'analyse des entités nommées, préconisent l'utilisation de corpus de presse.

Malgré les différents travaux de recherches effectués sur le traitement automatique de la langue amazighe, notamment sur la tâche de reconnaissance des entités nommées, la disponibilité des ressources linguistiques complètes et libres reste très limitée, d'ailleurs à notre connaissance, le seul corpus qui a été conçu pour la tâche REN amazighe contient d'environ 23 000 noms de places et 200 noms de personnes (Cieri, et al., 2008). Ce corpus n'est pas mis en ligne car il n'a pas encore été révisé et validé.

5. 1. 3 Propriété du corpus Amazighe « AmCorp »

Le recueil d'un corpus de documents html en langue Amazighe standard à partir d'Internet était une tâche moyennement difficile car le nombre de sites Web écrits en Amazighe sont très faibles.

La constitution du corpus amazighe (dorénavant, nous le nommerons AmCorp) s'est appuyée sur la collecte de quelques textes amazighes journalistiques déjà publiés sur le site web « mapamazighe[17]», le portail d'informations amazighes de l'Agence Maghreb Arabe Presse (MAP). Il est édité et destiné majoritairement aux locuteurs du Monde Amazighe. D'ailleurs les articles journalistiques s'imposent en bons candidats pour notre expérience. Ils représentent fidèlement l'écrit amazighe standard et ils sont accessibles à tous les locuteurs amazighes mis à part leur domaine de spécialité.

AmCorp contient l'actualité de toutes les régions du Maroc et du monde entier, les dernières dépêches, les activités régionales sur tout le territoire marocain notamment l'actualité sur les activités royales de SM le Roi Mohammed VI, les activités princières, les activités sportives, économiques, politiques et autres pour la période comprise entre mai 2013 et juillet 2015. Bref le choix des articles était sélectif pour couvrir les différents domaines et pour que la diversité, saisie en termes des entités nommées, soit moyennement grande. Ceci nous permettra d'appuyer les conclusions que nous en tirons.

Ce corpus compte :

- **900** articles journalistiques,
- Cumulant un total de **170.000**mots,

La Figure 5-1indique la répartition des entités nommées amazighes au sein du corpus « AmCorp ».

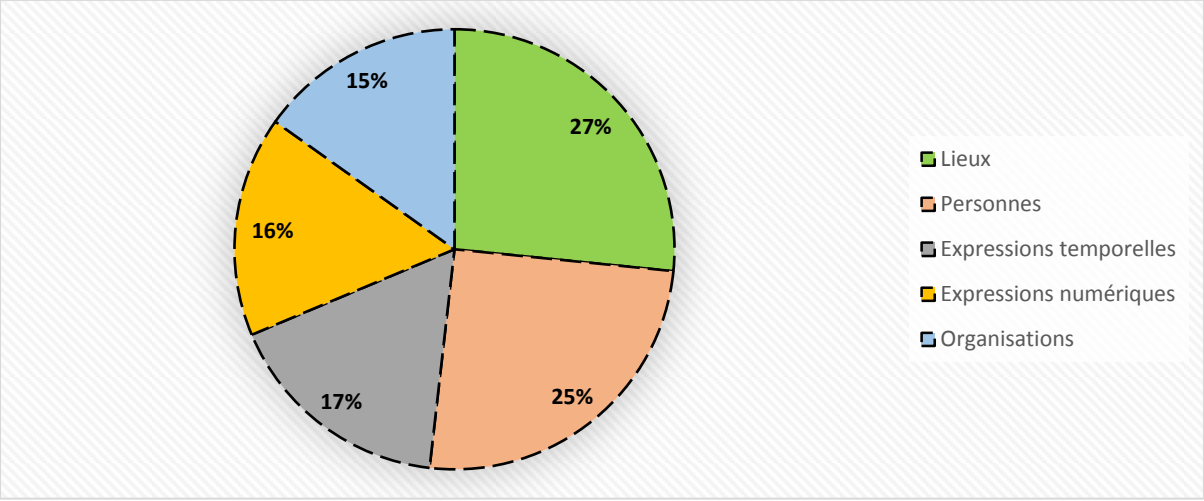

Figure 5-1 Répartition des types principaux d'entités

[17] http://www.mapamazighe.ma/am/

En termes de distribution, on constate qu'il y a une légère domination en termes de fréquence des entités nommées de type « lieux » qui sont suivis des « noms de personnes », la part restante (48%) étant répartie entre les 3 autres catégories qui sont les expressions temporelles (17%), les expressions numériques (16%), les organisations (15%).Il est à noter que les pourcentages sont très rarement utilisés dans notre corpus. Globalement, ce corpus est équilibré pour les types principaux d'entités.

Nous avons établi deux bases de données à partir du corpus initial AmCorp, une première d'entraînement qui sert à l'apprentissage, l'extraction et l'ajustement des règles d'extraction des entités nommées (**80%**), et une deuxième d'évaluation qui est utilisée pour évaluer les performances du système (**20%**).

5. 1. 4 Étapes de construction du corpus

Afin de rendre les données recueillies sur le Web exploitables, un certain nombre de traitements est nécessaire. Nous avons commencé par une première phase qui consiste à collecter les articles écrits en langue amazighe, notamment ceux édités et publiés dans le site officiel de la MAP amazighe. La deuxième phase sert au filtrage et à la normalisation du format des articles collectés, où nous avons effectué un « dé-balisage » des fichiers HTML et une conversion du format HTML en un format texte brut. Le schéma reproduit ci-dessous illustre le parcours des textes intégrés dans « AmCorp » (Figure 5-2) :

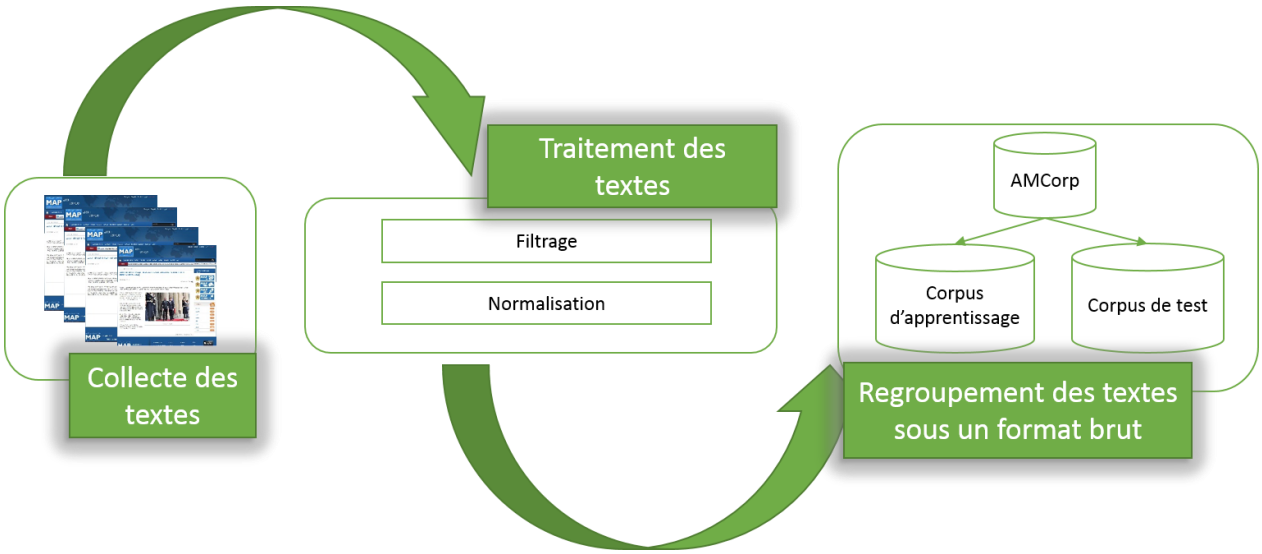

Figure 5-2 Processus de collection du corpus AMCorp

5. 1. 5 Segmentation du corpus

La procédure de segmentation d'une phrase en mots est plus ou moins facile pour la langue amazighe, similairement aux langues indo-européennes qui possèdent une écriture segmentée, les mots sont écrits avec espaces et il existe des ponctuations claires entre les phrases. Ces derniers définissent les modèles de caractères marquant la fin des phrases, ce qui nous permettra de procéder à la segmentation facilement du texte en des phrases et des phrases en token.

5. 1. 6 Annotation Manuelle du corpus

Une partie du corpus AmCorp a en effet été constitué et annoté manuellement, il s'agit d'un corpus annoté suivant un schéma d'annotation, en termes d'entités nommées, proche de celui de MUC 7. Dans ce corpus chaque entité nommée détectée est assignée à une catégorie convenable. Il est à noter que la typologie adoptée comme base pour notre travail comporte sept classes principales, dont voici la liste exhaustive (cf.

Tableau 5-1):

Entités Nommées	**Sous-types d'entités nommées**	**Exemples en amazighe**
Noms de Personnes (Person)	○ Prénoms, ○ Noms de famille	□□□□□□, Jawahir □□□□□□□, Lhamadi
Lieux (Location)	○ Continent, ○ Pays, ○ Ville, ○ Région, ○ Quartiers, ○ Boulevards, ○ Aéroports, ○ Stades	□□□□□□□, Chaouia □□□□□□, Khouribga
Noms d'Organisations (Organization) :	○ Écoles, ○ Organisations, ○ Ministères, ○ Instituts, ○ Entreprises, ○ Administrations, ○ Les équipes sportives,	□□□□□□□□ □□□□□□, Ministère de la Santé □□□□□ □□□□□□□□ □□□□□□ □□□□□□□, école

		nationale de commerce et de gestion
Expressions Temporelles (Date) :	○ Dates (Jour, Mois, Année), ○ Horaires, ○ Périodes, ○ Saisons	□□□□□, Mercredi □□□□□□□□, Décembre
Expressions Numériques y compris les Numéros (Numbers, Num) :	○ Numéros, ○ Mesures physiques (Poids, Volume, etc)	□□□□□□□□□□□, Mètre carré □□□ □ □□□□, dix-neuf
Expressions Monétaires (Money) :	○ Euro, ○ Dirham, ○ Dollars, …	□□□□□, Dollar □□□□, Dirham
Pourcentages (Percent) :	○ %, ○ pour cent…	%

Tableau 5-1 Les sept classes principales des entités nommées amazighes

L'annotation de ce corpus est accomplie de façon à fournir un ensemble d'informations nécessaires aidant dans la phase d'apprentissage. Il s'agit de décrire plus précisément les règles d'écriture sur lesquelles nos systèmes vont se reposer pour bien apprendre automatiquement les règles d'extraction des ENs et évaluer les résultats obtenus en les comparant à une analyse de référence.

Chaque élément annoté indique les positions de début et de fin des entités au sein du texte à l'aide d'une balise de format XML. Le type et le nom de l'entité dénotée sont également indiqués par des attributs, correspondant aux informations équivalentes spécifiées pour l'entité considérée dans la ressource.

La Figure 5-3présente ce format d'annotation pour un exemple d'entité nommée de type « Location, localisation » « Buznika, ⵣⵓⵣⵏⵉⴽⴰ » extraite du corpus AMCorp.

```
<Annotation Id="1100845" Type="Location" StartNode="132347" EndNode="132354">
</Annotation>
<Node id="132347"/>□□□□□□□<Node id="132354"/>. <Node id="132355"/>&#xd;
<Annotation Id="1249281" Type="Lookup" StartNode="132347" EndNode="132354">
<Feature>
<Name className="java.lang.String">minorType</Name>
<Value className="java.lang.String">country</Value>
</Feature>
<Feature>
<Name className="java.lang.String">majorType</Name>
<Value className="java.lang.String">location</Value>
</Feature>
</Annotation>
```

Figure 5-3 Exemple d'une entité nommée amazighe annotée sous GATE

5. 2 Expérimentation et Évaluation

5. 2. 1 Métriques de performance

Plusieurs métriques ont été définies pour évaluer les performances des systèmes de reconnaissance des entités nommées pour différentes langues, comme a été détaillé dans lasection 1. 3. 5. Dans notre travail, les performances pour la reconnaissance d'entités nommées amazighes sont évaluées en termes de rappel, de précision et de F-mesure utilisés dans la campagne MUC (Grishman, et al., 1996), étant donné le fait que la conférence d'évaluation MUC offre une référence intéressante pour mesurer l'efficacité des méthodes utilisées par les différents systèmes de reconnaissance des entités nommées. Rappelons ici la définition des différentes métriques :

$$Rappel = \frac{\text{Nombre d'entités correctement reconnues}}{\text{Nombre d'entités reconnues}} \quad \textbf{(5)}$$

$$Précision = \frac{\text{Nombre d'entités correctement reconnues}}{\text{Nombre d'entités dans le corpus}} \quad \textbf{(6)}$$

La F-mesure est la combinaison de la précision et du rappel et leur pondération. La formule de la F-mesure est :

$$F - Mesure = \frac{(1 + \beta^2) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}} \quad \textbf{(7)}$$

Classiquement Le paramètre β détermine lequel des paramètres (précision ou rappel) privilégié. Dans notre cas, nous voulons traiter également les deux mesures. Nous avons donc utilisé la valeur β=1. Nous obtenons alors la mesure appelée F1.

$$F - Mesure = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad \textbf{(8)}$$

5. 2. 2 Évaluation des systèmes de REN Amazighe sur notre corpus « AmCorp »

Tout d'abord, il est à noter que toutes les configurations testées n'apparaissent pas dans les résultats. Les seules qui ont été retenues c'est celles qui donnaient les meilleurs résultats sur la majorité des catégories. Rappelons que l'on vise à extraire des entités nommées amazighes (noms de personnes, noms de lieux, noms d'organisations, expressions numériques, expressions temporelles, expression monétaires, pourcentages) et à les typer. Rappelons aussi, qu'il ne s'agissait pas ici d'obtenir des meilleurs taux de reconnaissance mais de montrer qu'il est possible de créer des systèmes de reconnaissance des entités nommées pour la langue amazighe, une langue qui n'a pas suffisamment de ressources linguistiques en sens.

Système à base de règle « RENAM » :

La première évaluation consiste à évaluer les résultats obtenus par notre système à base de règles RENAM sur notre corpus « AMCorp ». Le Tableau 5-2illustre les valeurs du rappel, de la précision et de la F-mesure pour les différents types d'entités nommées.

Entités Nommées	Système à base de règles (RENAM)		
	Rappel (%)	Précision (%)	F-mesure (%)
Noms de Personnes	90	70	79
Noms de lieux	91	75	82
Noms des Organisations	71	85	77
Expressions numériques	97	91	94
Expressions Temporelles	94	82	88
Expressions Monétaires	64	84	73
Pourcentages	96	89	93

Tableau 5-2 Performances du système de REN Amazighe en utilisant une approche à base de règles

Étant donné que ce système représente la première expérience existante concernant l'extraction des entités nommées en amazighe, nous constatons que les résultats sont globalement assez encourageants. Nous pouvons aussi constater que la qualité de reconnaissance se trouve légèrement dégradée pour les différentes catégories.

Concernant la catégorie ENAMEX, les entités nommées de types « Lieux » ont une valeur du rappel égal à 91%, cette valeur est la meilleure par rapport aux autres valeurs du rappel pour les deux autres types d'entités nommées « Personnes » et « Organisations ». Ceci signifie que notre système de reconnaissance a pu localiser plus d'entités de type lieux. Pour les entités nommées de type « Organisation », la valeur du rappel trouvée est 71%, car il y a beaucoup d'entités « Organisation » n'ont pas pu être bien localisées. Cependant, la précision obtenue pour ce type d'entité nommée est élevée par rapport aux autres (85%) ce qui montre la fiabilité de notre système.

En ce qui concerne la deuxième catégorie « NUMEX », les entités nommées « Expressions numériques » ont une valeur de rappel (97%) et de précision (91%) assez élevée ce qui a influencé sur le résultat global de la F-Mesure (94%), ceci signifie que cette catégorie a été parfaitement détectable, et c'est, d'ailleurs, la catégorie qui a eu le plus grand score comparé aux autres.

En ce qui concerne la catégorie « TIMEX », la valeur f-mesure (88%) obtenue est assez bonne.

Système à base d'apprentissage :

Afin de permettre une comparaison objective des résultats produits par RENAM, nous avons conduit des évaluations en utilisant notre système à base d'apprentissage, nous présentons également les résultats obtenus sur le corpus de test dans le Tableau 5-3.

Entités Nommées	**Système à base d'apprentissage**		
	Rappel (%)	Précision (%)	F-mesure (%)
Noms de Personnes	85	56	68
Noms de lieux	94	77	85
Noms des Organisations	82	84	83
Expressions numériques	96	86	91
Expressions Temporelles	94	81	87
Expressions Monétaires	70	86	78
Pourcentages	100	100	100

Tableau 5-3 Performances du système de REN Amazighe en utilisant une approche à base d'apprentissage

En se basant sur les résultats illustrés dans le Tableau 5-3, nous avons constaté que la valeur de la F-Mesure est de 100% pour la catégorie des entités nommées « Pourcentages », ce qui montre que la configuration de ce système est suffisamment contrainte pour avoir un bruit quasi nul, mais ceci est dû à une cause principale liée à la faible occurrence des entités nommées de ce type, toutes les entités nommées de ce type ont été bien reconnue.

En revanche, la valeur de précision (56%) pour les entités nommées de types « Personnes » est la plus basse par rapport aux autres catégories, mais ce type d'entités a obtenu une valeur de rappel assez bonne, et bien évidement ceci a influencé le résultat de la f-mesure (68%).

Nous constatons aussi que par rapport au système précédent, il y a eu une légère dégradation de la qualité de reconnaissance pour les différents types d'entités, mais d'une façon plus marquante pour les entités nommées de types « Personnes », il y a eu une perte de (14%) en termes de Précision et de (11%) en termes de f-mesure. Ceci est dû à l'impact des caractéristiques choisies, cela démontre qu'il faut d'ajouter d'autres caractéristiques permettant d'offrir une amélioration significative à la performance de la reconnaissance.

Système hybride :

Le système hybride est la version finale de notre système. Nous renseignons dans le Tableau 5-4 les résultats produits par notre système hybride sur chacune des catégories des entités nommées.

Entités Nommées	**Système Hybride**		
	Rappel (%)	Précision (%)	F-mesure (%)
Noms de Personnes	76	89	82
Noms de lieux	88	97	92
Noms des Organisations	95	82	88
Expressions Numériques	95	97	96
Expressions Temporelles	96	89	93

Expressions Monétaires	73	88	83
Pourcentages	100	100	100

Tableau 5-4 Performances du système de REN Amazighe en utilisant une approche hybride

Les résultats obtenus du système hybride pour les différentes catégories prouvent une amélioration moyenne et qui peuvent être comparés favorablement avec les autres systèmes antérieurs. Les résultats montrent que dans tous les cas, il y avait une amélioration sensible par rapport aux résultats antérieurs. A la lumière de ces observations, nous pouvons conclure que, conformément à l'état de l'art et dans la lignée de travaux antérieurs dans différentes langues, la combinaison des deux systèmes a pu effectivement améliorer le taux de reconnaissance de toutes les entités comme est illustré dans la Figure 5-4.

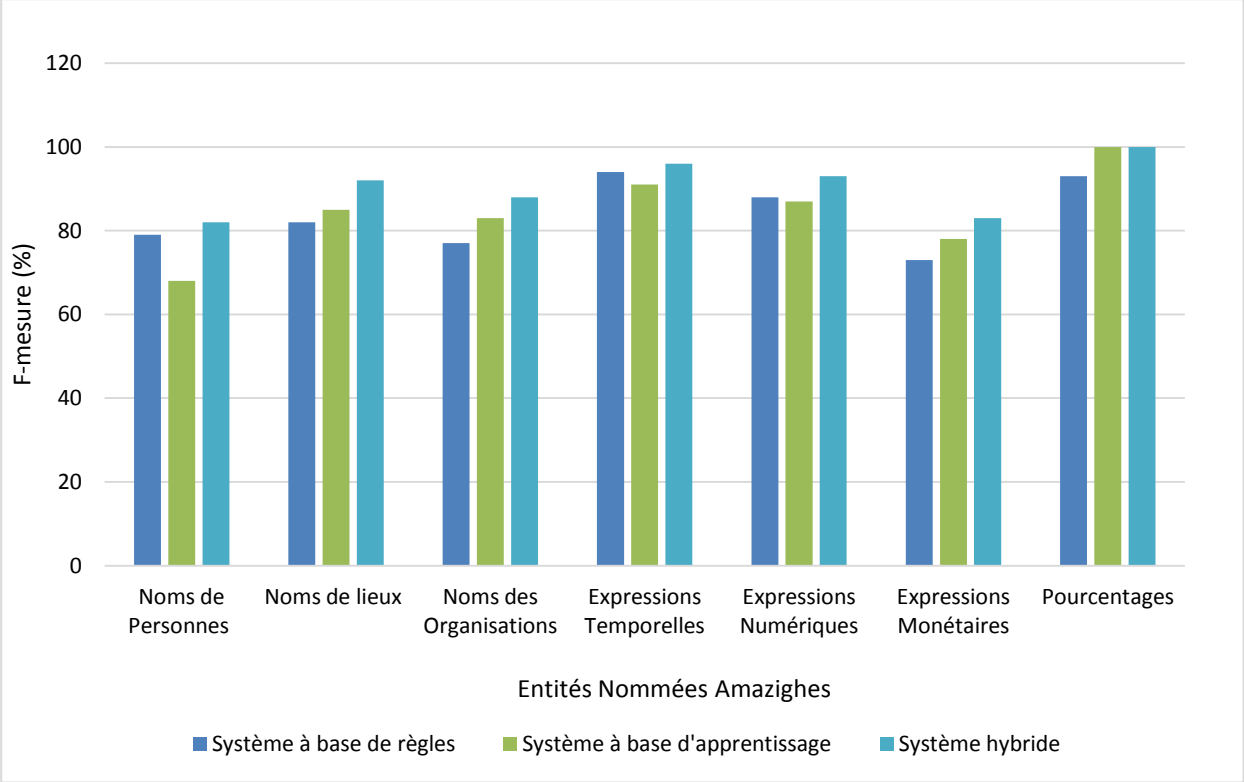

Figure 5-4 Comparaison des différents systèmes de la REN Amazighe sur le corpus d'évaluation

5. 3 Discussion

Le but de notre système est avant tout d'obtenir un score honorable, ceci est corroboré par les résultats que nous avons pu obtenir par ces différentes approches qui sont globalement satisfaisants. Néanmoins, nous avons constaté que certaines imprécisions persistent, il subsiste encore un nombre non négligeable d'erreurs d'annotation concernant la reconnaissance ou l'attribution d'une catégorie sémantique. A l'issue d'une analyse approfondie des résultats et des erreurs d'annotation obtenus au fur et à mesure de nos expérimentations tenues en utilisant les différents systèmes, nous pouvons tirer les limites de nos systèmes et les observations suivantes afin de donner également un aperçu du genre d'erreurs commises :

- **Erreurs orthographiques** : Tout d'abord, nous avons constaté que dans notre corpus peut figurer des graphies erronées liées au problème de translittération. En effet, dans notre cas, le corpus AMCorp contient beaucoup d'entités nommées de type personne de différentes origines et la majorité de ces ENs ont été translittérées, néanmoins, on a constaté que certaines entités n'ont pas été bien translittérées. Par exemple, le nom « □□□□□, [Karim] » est translittéré en amazighe à « □□□i□ », ou « □□□□□, [Muhnd] » au lieu de « □□□□□, [Muhmd] »conduisant les systèmes à ne pas déclencher nos ressources implémentées.
- **Erreurs typographiques :** Nous avons aussi constaté que de très nombreuses formes inconnues étaient liées à la typographie, sans que cela n'ait été détecté pendant la phase de nettoyage, comme par exemple l'oubli d'un espace entre les mots déclencheurs et les entités nommées en question (exemple : □□□□□□□□□□, [MassaHanan], □□□□ □□□□□, [MasaHanan]).
- Une partie des erreurs du système provient du fait que **certaines entités sont fortement ambigües**, « □□□□ □□ □□□□, Fkih ben saleh » qui désigne une entité de type lieux alors que nous avons défini une règle qui classe chaque entité qui s'écrit sous forme de « **mot+ bn + mot** » dans la catégorie Personne.
- Mis à part la limite précédente, nos systèmes ont été confrontés à d'autres types de problèmes liés à **la structure syntaxique des phrases**. En effet, la plupart des entités nommées de type organisation ont été correctement reconnues. Après une analyse nous avons constaté que la plupart des entités qui n'ont pas été bien reconnues ou partiellement reconnues n'existaient dans nos Gazetteers. Nous nous sommes

confrontés à un problème de délimitation des frontières de début et de fin de chaque entité, ceci est dû au fait que ces entités apparaissent généralement sous formes composées de plusieurs termes, ce qui rend leur délimitation plus difficile. La longueur du groupe de mot à prendre en compte est un problème difficile qui a généré beaucoup de reconnaissance partielle. Les entités nommées « organisations » présentes dans notre corpus contiennent presque toujours des combinaisons longues de mots auxquels nos ressources ne s'appliquent pas. Une analyse syntaxique plus profonde est nécessaire afin de résoudre ce genre de problèmes.

- **Manque d'information contextuelle :** En effet, sur les entités nommées mal reconnues, une grande majorité sont des EN pour lesquelles nous n'avons trouvé aucun mot déclencheur, aucun élément du contexte de gauche permettant de les catégoriser. Il nous faut donc trouver un autre moyen pour permettre la reconnaissance de ces entités nommées.
- **La forte polysémie** liée aux entités nommées pose un problème pour la mise en place de nos systèmes. En effet, nous sommes confrontés à plusieurs exemples comme : « tsga assa zag » qui signifie en premier temps « la région assa zag » et dans un deuxième temps « □□□□ □□□□□ » qui signifie « le jour de Mercredi », ('□□□…□…' (jar…d, entre… et…) qui est parfois un indicateur important indiquant la présence d'une entité nommée de type « Date » (Exemple : □□□ □□□□□□□□ □ 1876 □ 1884, jar isggasn n 1876 d 1884) ou indiquant la présence d'une entité nommée de type Lieu (Exemple : □□□ □□□ □ □□□□□, jar Fas d Mknas). Un autre exemple que nous avons pu détecter est que la règle qui identifie un nom de personne fondée sur la présence du déclencheur pertinent « □□□□□, Massa » a été déclenchée de manière erronée sur ce corpus (Exemple : (□□□□□ □□□□□□, Massa Fatima) ----- (□□□□□□□□□□, Massa Daraa). Faute d'avoir rencontré quelques occurrences minimes de ce type de construction sur notre corpus, il ne nous a pas été possible d'identifier et prendre en considération cette erreur lors de la construction de nos règles. Par conséquent, cette erreur a été corrigée et adaptée aux exemples présents dans notre corpus.
- **La couverture lexicale** : Bien que larges et détaillées, les ressources linguistiques que nous avons développées ne peuvent pas couvrir un large spectre des cas de figure des entités nommées qui évolue constamment, elle reste insuffisante. En d'autres termes, le système RENAM, basé sur une approche symbolique, a été conçu en prenant pour

exemple les données sur lesquelles il se destinait à être appliqué qui sont en premier degré les articles journalistiques. En conséquence, les règles et les ressources qu'on a construites se fondent majoritairement sur les propriétés des cas rencontrés dans notre corpus. En revanche, l'utilisation de RENAM sur un nouveau corpus de thèmes différents peut probablement générer du bruit, car les règles qui ont été construites n'ont pas été adaptées aux propriétés de ces nouvelles données.

5. 4 Conclusion

Dans ce chapitre, nous avons tout d'abord commencé par donner un aperçu sur la plateforme linguistique GATE et sur le plugin "NER Amazighe" que nous avons conçu. En effet, dans ce plugin nous avons construit toutes les ressources qui permettent de réussir la reconnaissance des EN amazighes.

Ensuite, nous avons présenté le corpus « AMCorp » que nous avons manuellement préparé et annoté. Nous avons aussi présenté une étude distributionnelle des différentes entités quipeuplent le corpus.

Enfin,différentes évaluations et expérimentations ont été conduites afin de mesurer les performances de notre travailet discerner ses limites en se basant sur les métriques d'évaluation pour la reconnaissance.

L'analysedes résultats de cette expérience nous a permis de mieux comprendre les raisons dela baisse des performances de nos systèmes par rapport à certaines catégories d'EN. Toutefois, les résultats expérimentaux obtenus restent prometteurs et semblent attester l'efficience de nos systèmeset qui peut être amélioré davantage ultérieurement.

Conclusion et Perspectives

6. 1 Conclusion générale

Dans ce travail de thèse, nous nous sommes principalement intéressés à la reconnaissance des entités nommées amazighes en particulier des noms de personnes, de lieux, des organisations, des expressions numériques, monétaires, les pourcentages et les expressions temporelles. Dans ce cadre, nous avons pu construire un corpus amazighe dédié à cette tâche et nous avons proposé trois systèmes de reconnaissance et catégorisation des entités nommées amazighes en se basant sur trois approches différentes qui ont contribué à améliorer les performances.

La première approche est à base de règles, elle utilise un ensemble de 22 lexiques et de 131 règles linguistiques permettant la reconnaissance des entités nommées amazighes. Elle a montré de bonnes performances et sa bonne efficacité. La deuxième approche est basée sur une méthode à base d'apprentissage en utilisant le classifieur SVM, sur la base des résultats de nos expériences, nous avons constaté que les performances des deux premiers systèmes sont très comparables. Ensuite, une approche hybride a été proposée en combinant les deux approches précédentes,une amélioration significativedu taux de reconnaissance a été observée.

Les systèmes proposés ont été réalisé à travers une étude typologique effectuée sur un corpus d'étude riche en entités nommées créé et annoté manuellement. En effet, il représente, à notre connaissance, le premier et unique version d'un corpus amazighe dédié à la tâche de reconnaissance des entités nommées amazighes, il contient un ensemble de 900 articles journalistiques comprenant différentes thématiques. Ce corpus nommée « AMCorp » nous a permis l'identification des lexiques et des grammaires nécessaires pour la reconnaissance des entités nommées en question. Il a aussi été construit dans le but de tester et valider nos systèmes et à servir d'autres chercheurs dans des domaines différents faisant usage à la tâche de REN, telles que les systèmes questions-réponses, les résumés automatiques et notamment la traduction automatique.

6. 2 Perspectives

Les perspectives envisageables pour l'amélioration et la poursuite de nos travaux sont nombreuses. Nous allons présenter par la suite celles que nous avons jugées pertinentes :

- Élargissement de la taille de notre corpus : l'adaptation de nos systèmes développés à de nouvelles tâches nécessite la disposition d'une taille importante de données.
- Extension et enrichissement de la structure de nos gazetteers afin de couvrir le maximum des entités nommées.
- Amélioration de la qualité des règles d'extraction d'entités nommées.
- Exploration de nouveaux descripteurs dans le module d'apprentissage : nous allons ajouter d'autres descripteurs qui améliorent les résultats pour certaines catégories dont le taux de reconnaissance est faible par rapport aux autres, comme le POS tagging, descripteurs morphologiques, etc.
- Combinaison du SVM avec d'autres classifieurs afin d'améliorer la reconnaissance.
- Généralisation de la REN pour d'autres contextes applicatifs, nous pensons particulièrement à la traduction automatique, tout en montrant que les entités nommées peuvent jouer un rôle très important dans la résolution de la tâche de traduction automatique et fournissent des solutions élégantes permettant d'alléger le processus d'analyse, en revanche, l'utilisation de la REN dans d'autres contextes applicatifs reste possible.
- L'intégration de notre plugin « NER Amazighe » dans une interface web afin de faciliter la tâche de reconnaissance pour d'autres utilisateurs.

Bibliographie

Abraham A, Rule Based Expert Systems, Handbook for Measurement Systems Design. - London : [s.n.], 2005. - pp. 909-919.

Al-Onaizan Y. et Knight K Translating named entities using monolingual and bilingual resources / éd. Linguistic Association for Computational. - [s.l.] : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Juillet 2002. - pp. 400-408.

Altun Y., Johnson M. et Hofmann T. Investigating loss functions and optimization methods for discriminative learning of label sequences. - [s.l.] : Proceedings of the 2003 conference on Empirical methods in natural language processing, Juillet 2003. - pp. 145-152.

Amrouch M. [et al.] Handwritten amazigh character recognition based on hidden Markov models. - [s.l.] : ICGST-GVIP Journal, 2010. - Vol. 10. - pp. 11-18.

Ananiadou S. et Mcnaught J. Text Mining for Biology and Biomedicine. - [s.l.] : Computational Linguistics, 2007. - Vol. 33. - pp. 135-140.

Appelt D.E. et Israel D.J. Introduction to Information Extraction Technology / éd. Press IOS. - [s.l.] : AI Communications Journal, 1999. - Vol. 12. - pp. 161-172.

Ataa Allah F. et Boulaknadel S. Amazigh Verb Conjugator. - Reykjavik : Proceedings of the 9th edition of the Language Resources and Evaluation Conference, 2014.

Ataa Allah F. et Boulaknadel S. Amazighe Search Engine: Tifinaghe Character Based Approach. - Las Vegas, Nevada : Proceeding of International Conference on Information and Knowledge Engineering, 2010. - pp. 255-259.

Ataa Allah F. et Boulaknadel S. Convertisseur pour la langue amazighe : script arabe - latin - tifinaghe. - [s.l.] : Actes du 2ème Symposium International sur le Traitement Automatique de la Culture Amazighe (SITACAM), 2011. - pp. 3-10.

Ataa Allah F. et Boulaknadel S. Light Morphology Processing for Amazigh Language. - Valletta, Malte : Actes de Language Resources and Human Language Technologies for Semitic Languages, 7th International Conference on Language Resources and Evaluation (LREC), 17-22 Mai 2010. - pp. 32-35.

Ataa Allah F. et Boulaknadel S. Pseudo-racinisation de la langue Amazighe. - Montréal : Proceeding of Traitement Automatique des Langues Naturelles., 2010.

Ataa Allah F. et Jaa H. Etiquetage morphosyntaxique: Outil d'assistance dédié à la langue Amazighe. - Agadir : Proceedings of the 1er Symposium international sur le traitement automatique de la culture Amazighe, 2009. - pp. 110- 119.

Ataa Allah F., Boulaknadel S. et Souifi H. Jeu d'étiquettes morphosyntaxiques de la langue amazighe. - [s.l.] : revue Asinag, 2014. - pp. 171-184.

Ataa Allah Fadoua et Boulaknadel Siham Amazigh Search Engine: Tifinaghe Character Based Approach. - USA : IKE 2010, 14-16 Juillet 2010.

Babych B. et Hartley A. Improving machine translation quality with automatic named entity recognition. - [s.l.] : Proceedings of EAMT/EACL 2003 Workshop on MT and Other Language Technology Tools, 2003. - pp. 1–8.

Babych Bogdan, Hartley, Anthony Improving Machine Translation Quality with Automatic Named Entity Recognition / éd. Linguistics Association for Computational. - Budapest : Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools: Resources and Tools for Building MT, 2003. - pp. 1-8.

Benajiba Y, Diab M et Rosso P Arabic Named Entity Recognition Using Optimized Feature Sets / éd. Linguistics Association for Computational. - Honolulu, Hawaii : Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, 2008. - pp. 284-293.

Benajiba Y. Arabic Named Entity Recognition [PhD thesis]. - [s.l.] : Techninal University of Valencia, 2009.

Benajiba Y. et Rosso P. ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. - [s.l.] : 3rd Indian International Conference on Artificial Intelligence (IICAI-07), Décembre 2007. - pp. 1814-1823.

Benajiba Y., Diab M. et Rosso P. Arabic named entity recognition: An svm-based approach. - [s.l.] : Proceedings of 2008 Arab International Conference on Information Technology (ACIT), 2008. - pp. 16-18.

Benajiba Y., Rosso P. et BenedíRuiz J.M. ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy / éd. Springer. - Berlin, Heidelberg : Gelbukh A. (eds). Computational Linguistics and Intelligent Text Processing. CICLing 2007. Lecture Notes in Computer Science, 2007. - Vol. 4394. - pp. 143-153.

Bender O., Och F. J. et Ney H. Maximum entropy models for named entity recognition / éd. Linguistics Association for Computational. - [s.l.] : Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Mai 2003. - Vol. 4. - pp. 148-151.

Berkhin P. et Becher J. Learning Simple Relations: Theory and Applications. - Arlington, VA : Proceedings of the 2nd SIAM ICDM, 2002. - pp. 420-436.

Bikel D. M. [et al.]Nymble: a high-performance learning name-finder / éd. Linguistics Association for Computational. - [s.l.] : Proceedings of the fifth conference on Applied natural language processing, Mars 1997. - pp. 194-201.

Bodenreider O. et Zweigenbaum P. Identifying proper names in parallel medical terminologies. - [s.l.] : Studies in health technology and informatics, 2000. - Vol. 77. - pp. 443–447.

Boukhris F. [et al.] La nouvelle grammaire de l'amazighe. - Rabat : IRCAM, 2008.

Boulaknadel S. Contributions au traitement automatique et l'apprentissage de l'amazighe médiatisé par la technologie [Mémoire d'HDR]. - [s.l.] : IRCAM, 2016.

Boulaknadel S. et Ataa Allah F. Building a standard Amazighe corpus / éd. Heidelberg Springer Berlin. - Prague : Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI), Aout 2011. - pp. 91-98.

Boulaknadel S. et Ataa Allah F. Online Amazighe Concordancer. - [s.l.] : Proceedings of International Symposium on Image Video Communications and Mobile Networks, 2010. - pp. 1-4.

Boulaknadel Siham, Talha Meryem et Aboutajdine Driss Amazighe Named Entity Recognition using a rule based approach. - Doha : AICSSA, 10-13 Nov 2014.

Brun C. et Hagege C. Intertwining deep syntactic processing and named entity detection / éd. Springer. - Berlin, Heidelberg : Advances in Natural Language Processing, 2004. - pp. 195-206.

Candillier Laurent Contextualisation, Visualisation et Evaluation en Apprentissage Non Supervisé. - lle III, Université Charles de Gaulle : [s.n.], 2006.

Chen H. H., Yang C. et Lin Y. Learning formulation and transformation rules for multilingual named entities. - [s.l.] : Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition, Juillet 2003. - Vol. 15. - pp. 1-8.

Chinchor N. et Robinson P. MUC-7 named entity task definition. - [s.l.] : Proceedings of the 7th Conference on Message Understanding, September 1997. - Vol. 29.

Cieri C. et Liberman M. 15 Years of Language Resource Creation and Sharing: a Progress Report on LDC Activities. In LREC. - Marrakech : Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), 28-30 Mai 2008.

Collins M. et Y. Singer. Unsupervised models for named entity classification / éd. ACL. - College Park, MD : Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999. - pp. 100-110.

Cucchiarelli Alessandro et Velardi Paola Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence / éd. Press MIT. - [s.l.] : Comput. Linguist.Journal, 2001. - Vol. 27. - pp. 123-131.

Cunningham H. [et al.] A framework and graphical development environment for robust NLP tools and applications. - [s.l.] : Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Juillet 2002. - pp. 168-175.

Cunningham H. GATE, a general architecture for text engineering. - [s.l.] : Computers and the Humanities, 2002. - Vol. 36. - pp. 223-254.

Cunningham H. Software Architecture for Language Engineering [PhD Thesis]. - [s.l.] : University of Sheffield, 2000.

Cunningham H., Maynard D. et Tablan V. JAPE: a Java annotation patterns engine. - [s.l.] : Department of Computer Science, University of Sheffield, Mai 1999.

Doddington G. R. [et al.]The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. - [s.l.] : LREC, May 2004. - Vol. 2.

Duan K-B et Keerthi S. S Which Is the Best Mutliclass SVM Method? An Empirical Study / éd. Springer. - Berlin : Sixth International Workshop on Multiple Classifier Systems, 2005. - pp. 278-285.

Ehrmann M. Les Entitées Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation . - Paris Diderot University : Doctoral dissertation, ., 2008.

Ekbal A. et Bandyopadhyay S. Bengali Named Entity Recognition using Support Vector Machine. - Hyderabad : Proceedings of the IJCNLP-08 Work shop on NER for South and South East Asian languages, Janvier 2008. - pp. 51–58.

EL Azrak N. et EL Hamdaoui A. Référentiel de la Terminologie Amazighe : Outil d'aide à l'aménagement linguistique. - Rabat : 4 ème atelier international sur l'amazighe et les TICs., 2011.

Elouahabi S., Atounti M. et Bellouki M. Amazigh Isolated-Word speech recognition system using Hidden Markov Model toolkit (HTK). - [s.l.] : International Conference on Information Technology for Organizations Development (IT4OD), IEEE, 2016. - pp. 1-7.

Es Saady Y. [et al.] Adaptation d'un correcteur orthographique existant à la langue Amazighe : cas du correcteur Hunspell. - [s.l.] : Actes du 1er symposium international sur le traitement automatique de la culture amazighe, 2009. - pp. 149-158.

Es Saady Y. [et al.]Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata. - [s.l.] : International Journal on Graphics, Vision and Image Processing, 2010. - Vol. 10. - pp. 1-8.

Fakir M., Bouikhalene B. et Moro K. Skeletonization methods evaluation for the recognition of printed tifinaghe characters. - [s.l.] : Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe, 2009. - pp. 33-47.

Ferrández O., Toral A. et Munoz R. Fine Tuning Features and Post-processing Rules to Improve Named Entity Recognition / éd. Springer. - Berlin, Heidelberg : International Conference on Application of Natural Language to Information Systems, Mai 2006. - pp. 176-185.

Florian R. [et al.] statistical model for multilingual entity detection and tracking. - [s.l.] : NAACL/HLT, 2004. - pp. 1-8.

Fong Y. S., Malançon B. R. et Wee A. Y. NERSIL- the Named-Entity Recognition System for Iban Language. - [s.l.] : Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, 2011.

Fourour N. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. - [s.l.] : Actes, Neuvieme Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN), 2002. - Vol. 1. - pp. 265-274.

Frain J., Fadoua A. A. et Ouguengay Y. A. Lexique amazighe pour mobile. - Rabat : Actes de la 6 ème conférence internationale sur les Technologies d'Information et de Communication pour l'AMazighe , Novembre 2014. - pp. 24-25.

Friburger N Reconnaissance automatique des noms propres, application à la classification automatique de textes journalistiques [Thèse de doctorat]. - Tours : Université François Rabelais, 2002.

Gaizauskas R. [et al.] GATE: an environment to support research and development in natural language engineering. - [s.l.] : Proceedings Eighth IEEE International Conference on Tools with Artificial Intelligence, Novembre 1996. - pp. 58-66.

Galliano S., Gravier G. et Chaubard L. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. - Brighton : Tenth Annual Conference of the International Speech Communication Association, 2009.

Galliano Sylvain, Gravier Guillaume et Chaubard Laura The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. - [s.l.] : Tenth Annual Conference of the International Speech Communication Association, 2009.

Greenberg Joseph Harold The Languages of Africa. - [s.l.] : Indiana University, 1963. - Vol. 25. - p. 171.

Grishman Ralph et Sundheim Beth Message Understanding Conference-6: A Brief History [Livre]. - Stroudsburg : Association for Computational Linguistics, Proceedings of the 16th Conference on Computational Linguistics, 1996. - Vol. Volume 1 : pp. 466-471.

Grouin C. [et al.] Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview [Livre]. - [s.l.] : Association for Computational Linguistics, Proceedings of the 5th Linguistic Annotation Workshop, 2011. - pp. 92-100.

Guo J. [et al.] Named entity recognition in query / éd. ACM. - [s.l.] : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval , Juillet 2009. - pp. 267-274.

Hartigan J. Clustering Algorithms / éd. Sons John Wiley &. - New York, NY : [s.n.], 1975.

Hewavitharana S. et Vogel S. Extracting parallel phrases from comparable data / éd. Springer. - Berlin, Heidelberg : Building and Using Comparable Corpora, 2013. - pp. 191-204.

IRCAM Conception et mise au point des polices tifinaghe. - [s.l.] : Centre des Etudes Informatiques, Systemes d'Information et Communication, plan d'action. [Online]. , 2003.

IRCAM Polices et Claviers UNICODE. - [s.l.] : Centre des Etudes Informatiques, Systemes d'Information et Communication. [Online], 2004.

Joachims T. Text categorization with support vector machines : Learning with many relevant features. - [s.l.] : ECML-98, Tenth European Conference on Machine Learning, 1998. - pp. 137–142.

Kaufman Leonard et Rousseeuw Peter J. Finding groups in data : an introduction to cluster analysis / éd. John Wiley & Sons Inc. - 2005. - p. 342.

Laabdelaoui R. [et al.] Manuel de conjugaison de l'Amazighe. - Rabat : IRCAM, 2012.

Lafferty J, McCallum A et Pereira F. C Conditional random fields : Probabilistic models for segmenting and labeling sequence data. - [s.l.] : ICML'01, Proceedings of the 18th International Conf. on Machine Learning, 2001. - pp. 282–289.

Loffler Laurian, A. M. La traduction automatique. - [s.l.] : Presses Univ. Septentrion, 1996.

Makhoul J. [et al.] Extraction, Performance Measures For Information. - [s.l.] : Proceedings of DARPA Broadcast News Workshop, Février 1999. - pp. 249-252.

McCallum A. et Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons / éd. Linguistics Association for Computational. - [s.l.] : Proceedings of the seventh conference on Natural language learning at HLT-NAACL, Mai 2003. - Vol. 4. - pp. 188-191.

McDonald D. Internal and external evidence in the identification and semantic categorization of proper names. - [s.l.] : Acquisition of Lexical Knowledge from Text Journal, 1993.

McDonald David D. Internal and external evidence in the identification and semantic categorization of proper names, Corpus processing for lexical acquisition. - Cambridge : MIT Press, 1996.

Mesfar S. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard [Thèse de doctorat]. - Besançon, Université de Franche-Comté : Thèse de doctorat, 2008.

Meur C. L., Gallinao S. et Geoffrois E. Conventions d'annotations en Entités Nommées. - 2004.

Miftah N., Ataa Allah F. et Taghbalout I. Sentence-aligned parallel corpus Amazigh-English. - Irbid : International Conference on Information and Communication Systems (ICICS), IEEE, 2017. - pp. 58-63.

Mikheev A., Grover C. et Moens M. Description of the LTG system used for MUC-7. - Fairfax, Virginia : Seventh Message Understanding Conference (MUC-7), 1998.

Nadeau David et Sekine Satoshi A survey of named entity recognition and classification. - [s.l.] : Lingvisticae Investigationes, 2007. - Vol. 30. - pp. 3-26.

Nédellec C., Nazarenko A. et Bossy R. Information extraction / éd. Verlag Springer. - [s.l.] : Handbook on Ontologies in Information Systems, chapter 31, 2009. - 2.

Nejme F.Z., Boulaknadel S. et Aboutajdine D. Développement de ressources pour la langue amazighe : Le Lexique Morphologique ElAmaLex. - [s.l.] : Actes de la conférence conjointe JEP-TALN-RECITAL, TALAF, 2016. - Vol. 11.

Nobata C. [et al.] Summarization system integrated with named entity tagging and IE pattern discovery. - [s.l.] : Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), 2002. - pp. 1742-1745.

Osenova P. et Kolkovska S. Combining the named-entity recognition task and NP chunking strategy for robust preprocessing. - [s.l.] : Proceedings of the Workshop on Treebanks and Linguistic Theories, Septembre 2002. - pp. 20-21.

Ouakrim O. Fonética y fonología del Bereber. Survey at the University of Autònoma de Barcelona. - 1995.

Oulhaj L. Grammaire du Tamazight. - [s.l.] : Centre Tarik ibn Zyad center for studies and research., 2000.

Outahajala M. [et al.] Using Confidence And Informativeness Criteria To Improve POS Tagging In Amazigh. - [s.l.] : Journal of Intelligence and Fuzzy Systems, 2014.

Paik W. [et al.] Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval / éd. Press MIT. - Cambridge : Corpus Processing for Lexical Acquisition, Language, Speech and Communications, chapitre 4, 1996.

Paliouras Georgio [et al.] Learning decision trees for named-entity recognition and classification. - [s.l.] : ECAI Workshop on Machine Learning for Information Extraction, 2000.

Paşca Marius [et al.] Names and similarities on the web: fact extraction in the fast lane. - Sydney : Proceeding ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 17-18 Juillet 2006. - pp. 809-816.

Pinto D. [et al.] Table extraction using conditional random fields / éd. ACM. - [s.l.] : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval , Juillet 2003. - pp. 235-242.

Piton O. et Maurel D. Les Noms Propres Géographiques et le Dictionnaire Prolintex, les lieux situés hors de France. - [s.l.] : INTEX pour la linguistique et le traitement automatique des langues, 2004. - p. 53.

Pizzato L.A., Molla D. et Paris C. Pseudo relevance feedback using named entities for question answering. - [s.l.] : Proceedings of the 2006 Australian Language Technology Workshop (ALTW-2006), 2006. - pp. 89–90.

Poibeau T "Deconstructing Harry" : une évaluation des systèmes de repérage d'entités nommées. - [s.l.] : Revue de la société d'électronique, d'électricité et de traitement de l'information, 2001. - 5. - pp. 25–33.

Poibeau T. Extraction Automatique d'Information. Du texte brut au web sémantique. - Hermès : [s.n.], 2003.

Poibeau T. Sur le statut référentiel des entités nommées. / éd. LAngues/LIMSI Association pour le Traitement Automatique des. - [s.l.] : Conférence Traitement Automatique des Langues 2005, 2005. - pp. 173-183.

Quinlan J. R. Machine learning: Easily understood decision rules. - [s.l.] : Computer Systems that Learn, eds. Weiss, SM and Kulikowski, CA, Morgan Kaufmann, 1991.

Rabi M., Amrouch M. et Mahani Z. Evaluation of features extraction and classification techniques for offline handwritten Tifinagh recognition. - [s.l.] : Global Journal of Computer Science and Technology, 2017.

Raiss H. et Cavalli-Sforza V. ANMorph: Amazigh nouns morphological analyzer. - [s.l.] : Press in Proceedings of the 5th Int. Conf. on Amazigh and ICT., 2012.

Rijsbergen C. J. Van Information Retrieval [Livre]. - Newton : Butterworth-Heinemann, 1979. - 2.

Rocktäschel T., Weidlich M. et Leser U. ChemSpot: a hybrid system for chemical named entity recognition. - [s.l.] : Bioinformatics, 2012. - Vol. 28. - pp. 1633-1640.

Sang K., et Tjong E. F. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. - 2002.

Satori H. et ElHaoussi F. Investigation Amazigh speech recognition using CMU tools. - [s.l.] : International Journal of Speech Technology, 2014. - Vol. 17. - pp. 235-243.

Sekine S. et Eriguchi Y. Japanese Named Entity Extraction Evaluation - Analysis of Results. - Saarbruecken : Proceedings of the International Conference on Computational Linguistices, 2000.

Sha F. et Pereira F. Shallow parsing with conditional random fields. - [s.l.] : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Mai 2003. - Vol. 1. - pp. 134-141.

Shaalan K. et Raza H. Arabic named entity recognition from diverse text types / éd. Springer. - Berlin, Heidelberg : Advances in Natural Language Processing, 2008. - pp. 440-451.

Shaalan K. et Raza H. NERA: Named entity recognition for Arabic. - [s.l.] : Journal of the Association for Information Science and Technology, 2009. - Vol. 60. - pp. 1652-1663.

Shaalan K. et Raza H. Person name entity recognition for Arabic / éd. Linguistics Association for Computational. - [s.l.] : Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Juin 2007. - pp. 17-24.

Silva C. et Ribeiro B. Inductive Inference for Large Scale Text Classification / éd. Springer. - [s.l.] : Kernel Approaches and Techniques, 2009. - Vol. 255.

Srikanth P. et Murthy K. N. Named Entity Recognition for Telegu. - Hyderabad : Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, Janvier 2008. - pp. 41–50.

Sutton C. et McCallum A. An introduction to conditional random fields. - [s.l.] : Foundations and Trends® in Machine Learning, 2012. - Vol. 4. - pp. 267-373.

Sutton C. et McCallum A. An introduction to conditional random fields for relational learning / éd. Press MIT. - [s.l.] : Introduction to statistical relational learning., 2006. - Vol. 2. - pp. 93-128.

Taghbalout I., Ataa Allah F. et El Marraki M. Pivot-based multilingual dictionary model for under-resourced languages. - [s.l.] : International Journal of Applied Engineering Research, 2017. - Vol. 12. - pp. 10342-10350.

Taghbalout I., Ataa Allah F. et El Marraki M. Towards UNL based machine translation for Moroccan Amazigh language.. - [s.l.] : International Journal of Computational Science and Engineering, 2016.

Takeuchi K. et Collier N. Use of support vector machines in extended named entity recognition / éd. Linguistics Association for Computational. - [s.l.] : proceedings of the 6th conference on Natural language learning, Aout 2002. - Vol. 20. - pp. 1-7.

Talha Meryem, Boulaknadel Siham et Aboutajdine Driss Development of Amazighe Named Entity Recognition System Using Hybrid Method. - [s.l.] : Research in Computing Science , 2015. - Vol. 90. - pp. 151-161.

Talha Meryem, Boulaknadel Siham et Aboutajdine Driss Enhancing performance of Hybrid Named Entity Recognition for Amazighe Language. - [s.l.] : Machine Learning Paradigms: Theory and Applications, 2018.

Talha Meryem, Boulaknadel Siham et Aboutajdine Driss NERAM : Named Entity Recognition for AMazighe language (RENAM: Système de Reconnaissance des Entités Nommées Amazighes) [in French]. - Marseille : TALN, 2014. - pp. 517-524.

Talha Meryem, Boulaknadel Siham et Aboutajdine Driss Performance Evaluation of SVM Based Amazighe Named Entity Recognition. - Cairo : [s.n.], 2018. - pp. 232-241.

Talha Meryem, Boulaknadel Siham et Aboutajdine Driss Système de reconnaissance des entités nommées Amazighes. - Paris : JADT 2014, 2014. - pp. 629-638..

Talha Meryem, Siham Boulaknadel et Driss Aboutajdine L'apport d'une approche symbolique pour le repérage des entités nommées en langue amazighe. - Luxembourg : EGC, 2015. - pp. 29-34.

Tjong Kim Sang E. F. et De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. - [s.l.] : Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Mai 2003. - Vol. 4. - pp. 142-147.

Vapnik V. N. The Nature of Statistical Learning Theory / éd. Springer-Verlag. - [s.l.] : Data mining and knowledge discovery, 1995.

Vapnik Vladimir, Guyon Isabel et Hastie Trevor Support vector machines. - [s.l.] : Mach. Learn, 1995. - Vol. 20. - pp. 273-297.

Vijayakrishna R. et Sobha L. Domain focused named entity recognizer for tamil using conditional random fields. - Hyderabad : Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, 2008. - pp. 59–66.

Xu Rui et Wunsch Donald C. Survey of clustering algorithms / éd. IEEE. - [s.l.] : IEEE Transactions on Neural Networks, Mai 2005. - 3. - Vol. 16. - pp. 645-678.

Zaghouani W. RENAR: A rule-based Arabic named entity recognition system. - [s.l.] : ACM Trans Asian Lang Inf Process (TALIP), 2012. - Vol. 11. - pp. 1-13.